CONCEIVABILITY ARGUMENTS

by

KATALIN BALOG

A Dissertation submitted to the

Graduate School-New Brunswick

Rutgers, The State University of New Jersey

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

Graduate Program in Philosophy

written under the direction of

Professor Brian Loar

and approved by

_____

_____

_____

_____

New Brunswick, New Jersey

October 1998

ABSTRACT OF THE DISSERTATION

Conceivability Arguments

by

KATALIN BALOG

Dissertation Director:

Professor Brian Loar

The dissertation addresses the mind-body problem, and in particular, the problem of how to fit phenomenal consciousness into the rest of reality. Phenomenal consciousness - the *what it's like* feature of experience - can appear to the scientifically inclined philosopher to be deeply mysterious. It is difficult to understand how the swirl of atoms in the void, the oscillation of field values, the firing of synapses, or anything physical can add up to the smells, tastes, feelings, moods, and so forth that comprise our phenomenal experience. There is a series of arguments - the so-called "Conceivability Arguments" - that spells out this puzzlement. If this arguments are successful then there is no place for phenomenal consciousness in a completely physical reality. The main conclusion of this dissertation is that the Conceivability Arguments are all dependent on a flawed premiss, and that

therefore these arguments - perhaps the most powerful among anti-physicalist arguments - all fail.

Conceivability Arguments begin with the premiss that we can *conceive* of *any* physical or functional facts obtaining without there being any phenomenal experience at all. This is sometimes expressed by saying that zombies (i.e., beings that are our physical and functional duplicates, but possess no phenomenal experiences) are conceivable. The claim that zombies are conceivable does not have to do with our powers of imagination, or our *psychological* abilities, but rather with the nature of physical and phenomenal *concepts.* The reason that zombies are claimed to be conceivable is that each person's thinking about phenomenal properties is completely dependent on her first person acquaintance with her own experience. 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.

I argue that these arguments all fail; they are refuted by a master argument that I call "the Zombie Refutation." 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 underlying the Conceivability Arguments - that principle that links conceivability and possibility - turns out to be self-refuting. Thus, the

zombies that the Conceivability Arguments supposedly demonstrate to be possible, return to undermine those very arguments; a fitting revenge.

Acknowledgments

This dissertation is the outcome of much discussion and collective thinking. I would like to thank Jerry Fodor, Brian Loar, Colin McGinn, and Ned Block for serving on my thesis committee, and helping me organize the thesis defense in the middle of July. I got helpful feedback and comments from all of them. My thesis supervisor, Brian Loar, has been especially influential on my work, both in terms of his own ideas which he has always generously shared with me, and in terms of his insightful criticism of the dissertation in progress.

Brian McLaughlin and Georges Rey spent abundant amounts of time in helping me getting clearer about what I want to say. Even though not officially on my committee, they acted as advisors in the best sense of the word.

I also thank John Biro, David Chalmers, Jennifer Church, Gary Gates, Joe Levine, Karen Neander, Jesse Prinz, Howard Robinson and Gene Witmer for helpful conversation and comments.

And last but not least my special gratitude goes to Barry Loewer who, both as my spouse, and as philosophical mentor, has been an invaluable

help in writing this dissertation. The dissertation would be very different without his input.

I dedicate the thesis to our as yet unborn son, Milan.

Table of Contents

> The feeling of an unbridgeable gulf between consciousness and rain-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)

INTRODUCTION

Phenomenal consciousness - the *what its like*[1] feature of experience - can appear to the scientifically inclined philosopher to be deeply mysterious. It is difficult to understand how the swirl of atoms in the void, the oscillation of field values, the firing of synapses, or anything physical can add up to the smells, tastes, feelings, moods, and so forth that comprise our phenomenal experience. One might be tempted to declare just on the basis of this thought that physicalism is false; that is, that it is false that every contingent fact, including those concerning phenomenal consciousness is, or is realized by, or is constituted by, physical facts. But it is one thing to declare that physicalism is false and quite another to argue that it is.

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

This dissertation concerns the most important arguments - the so-called “Conceivability Arguments” - for the claim that there is no place for phenomenal consciousness in a completely physical reality. Conceivability Arguments, which go back at least to Descartes (*Sixth Meditation*, in: Cottingham, Stoothoff, Murdoch 1984, Vol II, pp. 50-63), begin with the premiss that we can conceive of *any* physical or functional facts obtaining without there being any phenomenal experience at all.[2] This is sometimes expressed by saying that zombies (i.e., beings that are our physical and functional duplicates, but possess no phenomenal experiences) are *conceivable*.[3] 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.

The claim that zombies are *conceivable* does not have to do with our powers of imagination, or our *psychological* abilities, but rather with the

---

[2]Sometimes it is argued that the opposite is also conceivable, i.e., that it is conceivable that mental facts, especially experiences, occur without any physical or functional facts occurring. Cf. Descartes, ibid. It is not necessary, however, for the arguments under consideration, that conceivability should go both ways.

[3]I will use the term '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’s like' feature.

nature of physical and phenomenal *concepts.* The relevant notion of conceivability is this:

(Con) A statement S is conceivable, if it is logically consistent with the totality of conceptual truths, i.e., if -S is not a conceptual truth.

Conceptual truths (or analytic truths) are truths in virtue of meaning.[4] It is usually assumed that if S is conceivable then it is knowable *a priori* that S is conceivable. That is, it is assumed that someone who can entertain the thought that S, can come to know whether or not S is conceivable without empirical investigation. Failure to detect *a priori* any contradiction in S is 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 conceptual truths after all.[5]

The reason that zombies are claimed to be conceivable is that each person's thinking about phenomenal properties is completely dependent on

---

[4]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. For more on concepts in general, see Chapter One. On the question of the nature and existence of conceptual truths, see the discussion of two-dimensional semantics, and the Explanatory Gap Argument in Chapter Three.

[5]The claim that whether or not 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.

her first person acquaintance with her own experience. When I think *I have a headache*, I apply that concept to myself directly; not in virtue of the referent satisfying certain behavioral, physical, or functional characteristics. There are no conceptual connections between first person applications of the concept *headache* and physical, behavioral, or functional concepts.

Some philosophers have denied this: they claim that our *concepts* of various kinds of phenomenal states, e.g., our concept *pain*, are physical, functional, or behavioral concepts.[6] For example, a crude functionalist account of the concept *pain* is that it is the concept *an internal state typically produced by stimuli associated with harm which typically causes avoidance behavior*. Of course, if it is analytic that an internal state satisfying a certain functional specification is a pain, then zombies are impossible.

Others claim that, while concepts of kinds of experience, e.g., pain, nausea, etc. do not have functionalist analysis, the concept *conscious experience* does. For example, Shoemaker (1981) holds that zombies are conceptually impossible but inverted qualia is conceptually possible. This is an interesting view, but as we will see, this view will not block the Conceivability Arguments.

---

[6]See, for example Lewis (1966), Ryle (1949),White (1986), Levin (1986).

It seems to me that behaviorist and functionalist analyses of phenomenal concepts are quite implausible. When I think (same, I submit, for you) *I am in pain*, I am not thinking that I am behaving or disposed to behave in a certain 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 physical or functional concept. Whatever the ultimate nature of phenomenal experience, when I judge that I am having an experience of particular sort on the basis of having that experience, the concept I invoke is not a physical, behavioral, or functional concept. Rather, it seems to be a concept that I apply directly and spontaneously to the experience.[7]

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 infamous *private language argument*.[8] The argument relies on certain *a priori* considerations concerning the nature of meaning. The argument is quite

---

[7]Loar (1997) characterizes phenomenal concepts as "direct recognitional" concepts. I will discuss this view in Chapter Two.

[8]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 or not she is a zombie.

obscure, but the basic idea is that first-person direct uses of phenomenal concepts 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. But it would be an enormous understatement to say that there is no consensus as to exactly what the argument is, let alone whether it is sound. Current discussion of the conceivability arguments for the most part ignore it so I will as well. In the following, I will assume that there is nothing in our concept of consciousness that would allow us to rule out a priori the existence of zombies; zombies are conceivable.

From the premiss, considered a priori true, that zombies are conceivable, it is further argued that their existence is a *genuine metaphysical possibility.* This is a powerful result. If it is correct, and if, as I will assume throughout the dissertation, there are phenomenal facts, then physicalism is false. For it would mean that the totality of physical facts obtaining in our world, including the laws of physics, does not *necessitate* the phenomenal facts that obtain in our world.

Without further elaboration, the Conceivability Argument seems to commit a simple fallacy. On the face of it, the mere fact that it is conceptually possible for an **F** to exist without its being **G** does not entail that it is metaphysically possible for an **F** to exist without being **G**. After all, it seems

that we can conceive of water existing without being composed in part of hydrogen even though being composed in part of hydrogen is metaphysically necessary for being water. But during the past three decades, work on the semantics of modality and referring expressions (see, especially, Kripke 1972) has clarified the relationship between conceptual possibility and metaphysical possibility so as to take these objections into account.

This has lead to a revival of interest in Conceivability Arguments, and sophisticated versions of these arguments have been developed by Kripke (1972, pp. 144-155), Nagel (1974), Robinson (1993), Jackson (1982, 1993 and 1995, Lecture 2 and 3), Chalmers (1996, especially pp. 56-123), and others. Like their predecessors, these arguments rely on there being a link between conceivability and metaphysical possibility, but the formulation of this link now takes into account that conceivability does not *always* imply possibility. The proponents of these Conceivability Arguments claim that, while the conceivability of water not being H_2O fails to imply that it is *metaphysically* possible for water not to be H_2O, the conceivability of certain other statements, e.g., that there is a zombie world and that pain is not identical to C-fibre firing does imply their metaphysical possibility.[9]

[9]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

As we will see, the link between conceivability and possibility invoked by Conceivability Arguments entail that modal facts are ultimately reducible to facts about what is conceivable, and ordinary empirical facts 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.

There is a close cousin of the Conceivability Arguments thought up, but not endorsed, by Joe Levine (1983, 1993). This argument involves what Levine calls "the explanatory gap" between physical and phenomenal descriptions. Levine observes that, given a physical description of a person who is having certain experiences, we are completely left in the dark as to the phenomenal nature of those experiences. In other words, there appears to be an explanatory gap between the physical and the phenomenal descriptions. At one time there was also an explanatory gap between, for

---

statement for which conceivability implies possibility. The statement that a zombie world exists is supposed to fall under this kind.

example, ordinary talk of transmission of traits of parents to their offspring, and physical and biochemical descriptions. But that gap has mostly been bridged by genetic theory and molecular genetics. We, or at least molecular geneticists, have a pretty good understanding of how biochemical processes can provide the mechanisms that underlie the transmission of traits from parents to children.

Levine observes that the case of consciousness seems different. He argues that no current accounts bridge the gap, and that there are reasons to think the gap is in principle *unbridgeable*. The reason that the gap is unbridgeable is that we do not conceive of our own conscious states as satisfying some role - causal or otherwise -, but rather we grasp them directly. Because of this, knowledge of physical truths does not explain phenomenal truths. Now one can argue, although as I mention Levine seems agnostic about this, from the existence of this epistemological gap to the conclusion that there is an unbridgeable *metaphysical* gap between physical facts and phenomenal facts; i.e., that a zombie world is metaphysically possible.

The primary goal of this dissertation is to survey and evaluate Conceivability and Gap Arguments against physicalism. I aim to give them the strongest and most sympathetic formulations. But ultimately I will argue that they all fail; they are refuted by a master argument that I call “the

Zombie Refutation." The reason they fail has to do with the very nature of phenomenal concepts that gives rise to both the conceivability of zombies, and to the explanatory gap between the phenomenal and the physical. Because of the special nature of these concepts, the principle underlying the Conceivability Arguments - that principle that links conceivability and possibility - turns out to be self-refuting. Thus, the zombies that the Conceivability Arguments supposedly demonstrate to be possible, return to undermine those very arguments; a fitting revenge.

I will show that this special nature of phenomenal concepts explains the explanatory gap as well. That is, it is the nature of these concepts, i.e., the fact that we directly apply them to phenomenal properties that explains why no perspicuous physical explanation of phenomenal properties can be found. The explanatory gap is generated by the way we conceive of our phenomenal states; but it is conceptual in nature and is not indicative of any metaphysical gap. There is no need to suppose that physicalism is false in order to explain the explanatory gap; physicalism itself has the resources to do that.

The order of discussion is as follows: In Chapter One Physicalism is formulated and defended. Chapter Two discusses the nature of phenomenal consciousness and the concepts we apply to it. In Chapter Three I formulate the Conceivability Arguments due to Descartes, Nagel, White, Kripke,

Jackson, and Chalmers, and the argument suggested by the existence of the explanatory gap advocated by Levine. In Chapter Four I develop the Zombie Argument that refutes all the extant, and I believe every possible, Conceivability Arguments, as well as the Gap Argument, and defend it against objections. Chapter Four also further develops the account of phenomenal experience and phenomenal concepts proposed in Chapter Two that shows why the explanatory gap exists, and why most of us find Physicalism so incredible. The fact that Physicalism can explain why we find it incredible goes a long way toward disarming objections to it.

CHAPTER ONE:
PHYSICALISM

1.1 PRELIMINARIES

It will be useful to begin with a brief discussion of how I will be using certain words throughout this dissertation. The key words are “property”, “possible world”, “metaphysical necessity”, “concept”, and “conceptual necessity.”

Properties

By “properties” I will mean language independent entities that can be multiply instantiated. By calling them “language independent”, I mean that what properties there are is independent of whatever languages exist. However, properties are the semantic values, or the references, of predicates. A simple sentence, e.g., “Socrates is wise”, is true just in case the semantic value of “Socrates” instantiates the semantic value of “is wise”; i.e., the property **wisdom**.[10]

[10]Throughout, I will indicate that I am talking about properties with **bold face type**, and concepts with *italics*.

Some philosophers hold that not every predicate refers to a genuine property. On this view properties are sparse and only those predicates that satisfy certain further conditions, e.g., occur in the formulation of scientific laws, refer to properties. But I find it more useful to suppose that properties are abundant so that every meaningful predicate that can be used to express a truth evaluable thought refers to a property. We can call these abundant properties "common properties." Elite properties are ones that are special, e.g., are constituents of laws. On this usage "is grue" and "has negative charge" both refer to common properties but only the second is elite. Properties like **grue** are, relative to elite properties, like **has negative charge,** highly disjunctive properties of interest only to philosophers.

Following Lewis (1983) and Armstrong (1978) I will assume that some elite properties are fundamental (or perfectly natural), and others are constructs out of fundamental properties. Like them, I will not have much to say about what makes a property "fundamental", except to assume that the sciences are our best guide as to what the fundamental properties of our world are. Lewis and Armstrong also both agree that fundamental properties are *categorical*, that is, that they are individuated independently of laws and causal relations. This contrasts with the view that fundamental properties are individuated in terms of nomological connections (see Shoemaker 1979). On

this view laws, at least the fundamental ones, are expressed by necessary truths. I do not take a stand on this important issue here.

Non-fundamental properties are logical constructs out of fundamental properties. Among these are higher order or functional properties. A functional property F is a property that is instantiated by something x just in case x's appropriate parts, or x together with some other entities, instantiates properties P_1, P_2...P_n, and these instantiations are related to each other in certain specific ways; e.g., by causation. Clearly not every pair of properties, even fundamental properties, are co-instantiable by the same individual or individuals related to each other in certain ways. No particle, it appears, can have, for example, both positive and negative charge. I will say that property instantiations that can be co-instantiated are "compatible" instantiations.

Possible Worlds

A state of affairs is a collection or sum of compatible property instantiations. *Possible worlds* correspond to maximal states of affairs.[11] This account is purposely vague since I have not said anything about what makes

[11]The reader will note that basic individuals are absent from my possible world building. I am assuming that individuals are constructs out of property instantiations across possible worlds. This is certainly a controversial thesis, but will facilitate our discussion. Everything I will say about possible worlds is adaptable, with straightforward modifications, to the view that individuals are basic constituents of possible worlds .

a state of affairs *maximal* or what makes property instantiations *compatible.* Later we will look at an account of possibility, spelled out by Jackson (1995) and Chalmers (1996), that is based in conceivability and explicates the notion of compatible property instantiation in terms of conceivability.[12] For our discussion of physicalism this vague account will do.

Thoughts and statements

Thoughts[13] and *statements* possess truth values at possible worlds. A statement (or thought) S is *metaphysically necessary,* i.e., □S is true, iff S is true at every possible world. S is *metaphysically possible*, ◊S, iff S is true in at least one world. The notions of *property*, *possible world* and *metaphysical necessity* are interrelated. 'x is P' and 'x is Q' refer to the same property iff they are necessarily coextensive. That is, they refer to the same property iff the statement '(x)(x is P iff x is Q)' expresses a metaphysical necessity iff '(x)(x is P iff x is Q)' is true at every possible world.

---

[12] The account is neutral between views on which possible worlds are c*oncrete* (Lewis 1986a) and views on which they are *abstract.*

[13]On the general semantic outlook I adopt, the meaning of the linguistic entities (statements and terms) is derivative on, and can be accounted for, in terms of the meaning of the corresponding mental entities (thoughts and concepts). I will switch back and forth between the two, as the exposition requires.

Concepts

Concepts are mental representations. I will assume that there is a Language Of Thought (LOT), and that concepts can be thought of as "words" in this language. Concepts are constituents of thoughts. On the LOT hypothesis, thoughts are certain sentences in the LOT. They are what is believed, known, judged true, etc.

A concept, e.g., the concept *cat* refers to a property, **cathood**, by presenting it in a particular way. This is sometimes expressed by saying that concepts are or have *modes of presentation.* Concepts typically possess certain intrinsic properties (i.e., morphological features),[14] syntactic structure, conceptual role, and reference. Conceptual role is comprised of the inferential dispositions involving thoughts containing that concept, as well as causal relations connecting the concept to other intentional states, including phenomenal states. Conceptual role is idiosyncratic in that it is immensely likely that no two thinkers will have exactly the same inferential dispositions.

[14]For words in LOT these intrinsic features can be thought of as being like the spelling of a word in a natural language. (By saying this I do not want to beg the question against Dualism; the intrinsic features in question might be non-physical.) And there may be other kinds of representations, e.g., images that possess other kinds of intrinsic features.

It is a controversial question which aspect of concepts (some intrinsic property of concepts, syntactic structure, conceptual role, or a combination of the above) is or determines its mode of presentation. Frege thought that *sense*, i.e., *mode of presentation*, is some primitive, possibly non-physical property of concepts that we directly grasp. Fodor (see, e.g., Fodor 1997) identifies the mode of presentation of concepts with morphological and syntactic features. Specifically, Fodor holds that conceptual role plays no role in individuating concepts (and so is no determinant of mode of presentation).[15] Others, however (see, e.g., Block 1986, and Peacocke 1992), accept that the conceptual role of a concept is a determinant of its mode of presentation. Proponents of this view divide among those (see Block 1986) who think that the totality of a concept's inferential role is individuative, and those (see Peacocke 1992) who hold that some inferential relations are special. These *special* inferential roles are thought to be meaning constituting, i.e., part of the concept's mode of presentation, while the rest of a concept's inferential roles are irrelevant to mode of presentation.

It is also usually thought by proponents of this view that it is a priori knowable to a thinker who possesses the relevant concepts *exactly which* inferential roles are meaning constituting. Underlying this assumption is the

---

[15]Fodor excepts logical concepts *and*, *or, etc.* For these concepts conceptual role is individuative.

idea that the meaning, or mode of presentation, of one's concepts is *a priori* available. For example, a thinker may be inclined to infer from 'x is a cat' both 'x is an animal' and 'x likes liver', but the thinker will know that only the former is individuative of the concept *cat*. If it is further assumed, as proponents of this view typically do, that such inferences are truth preserving, then this will make the inference 'x is a cat' therefore 'x is an animal' *analytic*, i.e., meaning constituting and a priori.

I will say that a thought that is true in virtue of the meaning constituting inferential roles of its constituent concepts is a conceptual truth. A thought that is logically compatible with all conceptual truths is conceptually possible, or, for short, *conceivable*.

I myself will remain agnostic in this dissertation on the correct theory of concepts, and specifically on the issue of whether or not inferential role plays a large part in individuating concepts. But later I will argue that there are certain features of the role of *phenomenal concepts* that are what makes these concepts phenomenal. Moreover, since the Conceivability Arguments are arguably all committed to conceptual role being at least partly individuative of concepts, and to there being many conceptual truths, for the sake of the argument I, too, will take that for granted for large parts of the dissertation.

The distinction between *properties* and *concepts* will be very important to my discussion. Properties are in the world, while concepts are in the mind (which is a small part of the world). Distinct concepts, however these are individuated, may refer to the same property. For example, the concepts *triangular* and *trilateral* refer to the same property since it is metaphysically necessary that whatever instantiates one instantiates the other. In this case the thought that all and only triangular things are trilateral things is both metaphysically and conceptually necessary. Its conceptual necessity derives from the fact that the inferential roles of the two concepts alone determine that the concepts corefer. But this is not always the case. For example, the concepts *water* and H_2O refer to the same property, even though the thought that *Water is* H_2O is not conceptually true. We can conceive of one and the same property via two very different modes of presentation. Exactly how different concepts can be and still corefer is an interesting issue in general. But clearly they can be quite different. Thus scientific concepts like *molecular motion* and everyday folk concepts like *heat* can corefer. Some concepts involve indexical modes of presentation, e.g., *that kind of plant*. Such concepts and non-indexical scientific concepts can also corefer.

1.2 FORMULATING PHYSICALISM

Physicalism is a metaphysical view of the basic constitution of the universe. It is variously expressed as the view that the world is nothing but, or nothing over and above the physical world; that a completed physics (presumably an improved version of present-day physics) will give us a comprehensive and correct theory of the universe; that the complete story of the world is the physical story;[16] or, more colorfully, that all God had to do to create our world is to create the distribution of fundamental physical quantities in space/time and to create the laws of physics; all true propositions are true in virtue of these. A formulation of physicalism purged of theological references is due to David Lewis (1983)[17]:

> P*. Among worlds where no natural property *alien*[18] to the actual world is instantiated, no two differ without differing physically; any two such worlds that are exactly alike physically are duplicates simpliciter.

---

[16]That, of course, does not mean that all our language can be finitely translated into the language of physics.

[17]To help the reader I included two appendices for easy reference to definitions and arguments appearing in the text.

[18]A property, according to Lewis, is *alien* to a world iff it is not analyzable as a Boolean construct out of natural properties all of which are nomologically possible to be instantiated by inhabitants of that world.

P* is a substantial metaphysical claim, since it privileges physical properties as fundamental. Another obvious virtue of P* is that it is contingent, since there are worlds at which it is false. We want a formulation of physicalism to allow, for example, worlds with ghosts, where physicalism fails. If the actual world contained instantiations of ectoplasmic properties, then physicalism would be false. P* gets this right: the actual world would have a physical duplicate that is not a duplicate *simpliciter*, since it would fail to contain the ectoplasmic individuals.

The reference to *alien* properties serves to rule out worlds which are physical duplicates of the actual world but contain extra non-physical properties from the range of relevant possible worlds. If the restriction was dropped from P*, then, even if only physical properties were instantiated at a world **w**, the mere metaphysical possibility of worlds that are physically just like **w**, but contain some extra, say, ectoplasmic[19] individuals, would render physicalism false at **w**. A world which is exactly like **w** all physical respects but where there are ectoplasmic entities, would differ from **w** without differing physically, and it would not be a duplicate of **w**; therefore P* would not capture physicalism if the reference to alien properties were dropped.[20]

---

[19]'Being ectoplasmic' is a stand-in to refer to a non-physical fundamental property.

[20]Horgan (1982) has a definition similar to Lewis's; it also relies on the

Physicalism appears to be a strong and possibly, but not obviously, true (or false) claim about the nature of our world. To the contrary, Crane and Mellor (1990) have argued that "there is no question of physicalism", since it is either obviously false or trivially true. They claim that if "physical" in P* means "properties and laws expressed by current physical theory" then P* is very likely false since it is very likely that current physics does not provide a complete or completely correct inventory of physical properties and laws.[21] No complete description of a world in the language of current physics is plausibly a description of the actual world.

The natural response to this problem is to appeal to *idealized physics*. The ideal physical theory is the theory that correctly describes the structure of space/time, the motions of all macroscopic objects and the laws governing these motions, as well as the laws governing whatever other entities and properties exist. Crane and Mellor claim that this characterization renders physicalism trivially true, since it would obtain as long as there is an

---

notion of "alien" property. Jackson (1994) and Chalmers (1996) also gives a formulation along these lines; instead of *alien properties*, however, Jackson's formulation uses the notion of a *minimal physical duplicate*, and Chalmers talks about *copies* of a world. It does not make much difference from the point of view of the discussion at hand which definition we use; I will stick to Lewis's formulation.

[21]It is generally thought that current physical theory cannot be the last word since there are conflicts between quantum theory and both special and general relativity theory.

ideal physical theory even if that theory contains intentional and phenomenal predicates. Of course, such an idealized theory is not what physicalists have in mind. If it turns out that to account for paradigm physical events (motions of particles, changes in field values, etc.) mental predicates *must* be employed, then physicalism is false.

One way of responding to this worry is to produce a positive characterization of "physical predicate." In current physics, atomic predicates and function expressions are micro-physical. They apply to points of some appropriate space/time structure. For example, fields are specified by functions assigning values to points of space/time. Mental predicates are certainly not like this. But it would be rash to assume that anything that would count as physics must be like this. Fortunately, for the purposes of discussing the conceivability arguments, a *negative* characterization of the language of ideal physics will do. We should simply require that the ideal physical language not contain any mentalistic predicates as atomic predicates.[22] Fodor makes the point effectively:

> I suppose that sooner or later the physicists will complete the catalogue they have been compiling of the ultimate and irreducible

[22]Some physicists (see, e.g., Wigner 1967) have proposed that, to account for the supposed "collapse" of the wave function when a system is measured, appeal must be made to consciousness or the intentions of the measurer. If so, then the ideal theory would not be a physical theory and physicalism would be false.

> properties of things. When they do, the likes of *spin*, *charm,* and *charge* will perhaps appear on their list. But *aboutness* surely won't; intentionality simply doesn't go that deep. (1987, p. 96)

I would only add that phenomenal consciousness properties won't be on the physicists list either.

We can then formulate a successor to P* as follows:

> P Among worlds where no property alien to the actual world is instantiated, any two that are exactly alike with respect to their complete descriptions (including specification of the fundamental laws) in the language of the ideal fundamental physical theory are duplicates simpliciter.

In P, none of the atomic predicates of the language of fundamental physics are intentional or phenomenal predicates.[23] The descriptions referred to in P will be infinite, specifying perhaps the values of all fields at every point of space/time. The full physical description also includes a specification of the laws of physics, stating that they are the laws. This latter statement may, depending on the correct account of laws, not be conceptually entailed by the rest of the full physical description.[24]

---

[23]This condition does not preclude there being logically complex predicates which are intentional or phenomenal predicates in the language of ideal physics. Making P work requires characterizations of intentional and phenomenal predicates. I do not offer that here, but in Chapter Two I discuss the nature of phenomenal predicates.

[24]On Humean accounts of laws (Lewis 1986b, Loewer 1997), the distribution of non-nomic properties conceptually implies which generalizations are laws. In contrast, on non-Humean accounts (Armstrong 1980, Dretske 1977), what laws there are is not so implied.

1.3 SUPERVENIENCE PRINCIPLES

If K* is a statement in ideal physics, and T is a truth (not necessarily in the language of physics), then truths of the form $\Box(K^* \supset T)$ are called "physical supervenience principles". They specify how non-physical statements supervene on physical statements. There is an important consequence of P that we will need later. Suppose that T is some truth about the actual world, and K is the *complete*, true description of the actual world in the language of ideal physics. Then one might think that if P is true then $\Box(K \supset T)$ is true, since worlds just like the actual world physically (worlds at which K holds) must be worlds where T is also true. In other words, if physicalism is true, every actual truth supervenes on the complete physical description. Typically truths will supervene on much less than the entire physical description. So, for example, that there is an ice cube in a certain region R of space/time will supervene on fundamental physical descriptions of that region. Physical facts concerning what is going on outside of R are metaphysically irrelevant to whether or not there is an ice cube in R. Similarly, it is plausible that statements about a person's phenomenal experiences supervene on her neurophysiology. What is going on outside her brain is metaphysically irrelevant.[25]

[25]It is now widely accepted that intentional content often does not supervene on neurophysiology. Theories that hold that phenomenal experience is identical to, or partly constituted by, intentional content may reject this supervenience claim.

Interestingly, even though P is true, there are some statements for which $\Box(K \supset T)$ fails. It fails for those statements that, intuitively, depend on the totality of the distribution of fundamental properties. I will call such statements "global statements". An example is 'there are exactly n electrons in the history of the universe'. K may metaphysically imply that there are at least n electrons in the history of the universe, but it will not imply that there are exactly n of them. We need to add a conjunct saying that K is the *complete* physical description of the universe. Call this statement C.

There is another difficulty. Suppose that Q expresses a functional property that can be realized non-physically, and suppose further that the statement 'There are n Qs' is true. Then, even if P is true, $\Box(K \supset$ there are n Qs) will not be true, since there are worlds at which K is true which contain extra physical realizations of Q. For this reason we need to add another conjunct to K saying that all fundamental properties are physical properties.[26] Call this statement F. With these amendment we obtain

(E) For any true statement T,
$\Box(K \& C \& F \supset T)$,

which is equivalent to P.

If P is true then, for most true statements T, $\Box(K \supset T)$ will be true. Those truths for which this does not hold are, as we mentioned, "global"

[26]This corresponds to the stipulation about alien properties in P.

truths, in that their truth depends on the global distribution of fundamental properties. It is clear that positive phenomenal properties, e.g., being in pain, are not global.

Most physicalists would deem P too weak to fully express physicalism. It is compatible with there being anomalous physical events, and even with there being emergent laws involving higher level properties. Further, as Kim (1990) observes, it does not succeed in capturing the idea that all facts obtain *in virtue of* physical facts. For example, P could be true in virtue of some strange set of "quizzical" properties that underlie both physical and non-physical property instantiations (see Witmer 1997, p. 137). But for our purposes, we can stay with P. Any version of physicalism worthy of the name entails P; and it is by trying to refute P that the Conceivability Arguments are trying to refute physicalism. It will be enough then to show that the Conceivability Arguments have not succeeded in this to show that they have not refuted physicalism. Further, in the next section I will say a bit more about how physicalism can capture the idea that mental descriptions are true *in virtue of* physical facts.

1.4 REDUCTIONISM

P entails that every truth is metaphysically implied by truths in the language of physics.[27] But it does not entail that languages or conceptual systems other than that of fundamental physics are dispensable. There are many different levels of description, many different conceptual systems, that we employ in thinking about the world. Some concepts - those that conceive of middle size objects in terms of typical shapes, colors, etc. - are easy for humans to correctly apply on the basis of perception; e.g., *is a tree, is a rock*, etc. Other concepts may require specialized instrumentation or theories to apply, e.g., *is a virus.* It is not consequence of P that these concepts are either definable in terms of physical concepts or fail to pick out genuine properties. Statements in higher level vocabularies are typically more salient to us than statements in the language of ideal physics. Indeed, even those who have the requisite concepts to understand statements in ideal physics will not be in a position to employ those statements in ordinary discourse. Further, our understanding, to the extent we have it, of physical concepts plausibly presupposes the possession of higher level concepts.

---

[27]That is, if we set aside the difficulties mentioned in the last section involving global properties. From now on I take it for granted that this is how we proceed.

As far as I can see, all this is compatible with P. It is also compatible with P that there are true counterfactuals and laws couched in higher level vocabularies and that higher level statements can be confirmed. Thus it is compatible with P for there to be *special sciences* that develop more or less autonomously from physics. All that is required by P is that any truth including truths involving laws, counterfactuals, confirmation, etc. is metaphysically entailed by statements of ideal physics.[28]

What is the status of supervenience principles? Some philosophers have observed that they should not be thought of as expressing *brute facts*. To treat them as such would be to make supervenience an entirely mysterious relation. One would like the supervenience relations required by P to themselves be explained. As Terry Horgan[29] says

> unless psychophysical supervenience facts are themselves explainable, the instantiation of mental properties is not explainable on the basis of physico-chemical facts, but only on the basis of such facts *plus metaphysically fundamental inter-level, supervenience facts*. (p. 478)

Now, in fact, there are explanations of some physical supervenience principles and it is generally thought that only complexity stands in the way of

---

[28]P places some restrictions on the special sciences, since the laws and causal processes posited in the special sciences must be implementable by physical causal processes.

[29]See the entry "Physicalism" in Guttenplan (1994).

producing explanations of many others. These explanations involve *perspicuous reductions* of higher level predicates (and concepts) to lower level and, in some cases, physical predicates. I take it that the reductive relation relates concepts, and not the properties they refer to, since it makes no sense to speak of a property being reduced to itself.

A paradigm case of perspicuous reduction is the reduction of *water* to *H2O*. According to the usual account, our concept *water* is the concept of that actual substance that happens to be a clear liquid, quenches thirst, fills the oceans, etc. It is an empirical fact that the compound H_2O (the substance referred to with the concept H_2O) satisfies this specification. This justifies the statement that water is identical to H_2O. Given this reduction, we are in a position to explain a supervenience principle of the form

$\Box(...H_2O...\supset...water...)$.

I call such an explanation 'perspicuous reduction' since, once the explanation is given, it is perspicuously clear that the proposed supervenience claim is true; it would not make sense to doubt its truth any more.

Functionalism provides another example. We can see how the facts expressed in the higher level discourse might supervene on the facts expressed in the lower level discourse if we see that all there is to the satisfaction of higher level predicates is the playing of a certain causal role. Any lower level description that displays the appropriate causal relations

between the properties referred to, will thereby be a reductive account of the higher level description. Successful functional reductions abound in, e.g., the biological sciences; an example is the reduction of genes to configurations of DNA-molecules.

If, e.g., analytic functionalism about the mental was true, then we could see how the mental is reducible to the physical. Analytic functionalism is the view that mental predicates can be specified as referring to properties having a certain causal role *vis a vis* sensory inputs, behavioral outputs, and other mental states.[30] Any neuro-physiological description of a brain-state that satisfied that causal role would then serve as a reductive account of the mental state in question. But it is a wide-spread view that analytic functionalism is not an adequate account of mental properties, at least as far as phenomenal properties go. More on this later.

---

[30]A physical property, e.g., realizes a functional property iff it plays the causal role specified in the functional definition. The majority view is that psychological properties, like higher level properties in general, are multiply realized. For a classic treatment of the subject see Fodor's "Special Sciences, or The Disunity of Science as a Working Hypothesis", in Fodor (1975), pp. 9-25.

This, of course, does not require that every *metaphysically* possible instance of a functional property would be realized by an instance of a physical property. Functionalism allows not only for multiple physical realization, but also for worlds where, e.g., mental properties are realized by non-physical properties, or even where mental properties are basic.

Functionalism about certain higher level predicates allows for the failure of Physicalism, since functional properties can be realized by non-physical properties as well as physical properties. But it is also compatible with it, and the physicalist's bet is that in all the nomologically possible worlds functional properties are always realized physically.

Property identity, and functionalism are both examples of reductive accounts of higher level predicates. *'Reductive'* is used here in a revisionist sense. Classical reductionism demanded identity between the properties referred to by higher level and lower level predicates. For our purposes, however, it will be useful to introduce the concept of reduction broadly so that a functional account of a higher level predicate will also count as a reductive account, since it can be easily seen how a functional property can be realized by lower level (ultimately, physical) properties. I want to leave the question open whether property identity and functionalism (or perhaps structuralism) exhaust the possible varieties of a reductive account, or there are hitherto unthought-of ways in which lower level properties might be able to realize higher level properties.

An important question is whether all supervenience principles can be given a perspicuous reductive explanation. This is a question that I will return to when I consider the Conceivability Arguments in Chapter Three and Four. For now, I want to argue that even if there is no perspicuous reductive

explanation of a supervenience principle, it does not mean that the principle expresses a brute fact.

To say that a statement expresses a brute fact is to say that what makes that statement true cannot be accounted for in terms of the properties and individuals that constitute the fact. The reduction of *water* to H_2O was based on the observation that the concept *water* is associated with a contingent reference fixing description 'the substance that is a clear liquid, quenches thirst..etc.' But it may be that a concept refers to a property but not via a descriptive mode of presentation like this. Suppose, for example that concepts *C* and *D* are not associated with contingent reference fixing modes of presentation but refer directly to the same property, **E**. I am not now arguing that this is a possibility, but only that if it is then the fact expressed by *C is D* is perfectly un-mysterious, even though it cannot be explained in the way we explained *Water is H2O*. A similar point applies if *C* refers to a functional property **C** and *D* refers to a property **D** that realizes **C**, and *C* and *D* directly refer. $Dx \supset Cx$ will then be true, and the fact it expresses will be entirely un-mysterious. It is just a realization fact and as such is necessary.

In fact some hold that functionalism is true about mental properties, even though mental *predicates* cannot be analyzed in functional terms. Functionalism as a *metaphysical* doctrine says that what *makes* a state a particular mental state is just the fact that that state is of a type such that it is

related to sensory input, other mental states and behavior in some specified way. A metaphysical functionalist does not have to be an analytic functionalist as well; she can have almost any view about the semantics of mental terms as long as she maintains that the nature of the phenomena referred to by those terms is functional.[31]

If these cases are possible, we can say that *metaphysical* reductionism holds between the higher and lower level predicates. Even though the semantics of the predicates, together with contingent facts, does not reveal the identity relation or the functional realization relation holding between the properties referred to, identity or the realization relation can still hold. So a reason to think that some form of the identity theory or functionalism holds in the metaphysical sense would be a reason to think that there is an explanation for the supervenience facts.

P entails supervenience principles. But it does not by itself entail that supervenience principles can be reductively explained or that they are grounded in the natures of properties. For all P says these principles might be brute. If we were forced into such a view then I think that would be reason to doubt P. But I will argue that there is no reason to think that there are true

[31]See, e.g., Putnam (1967), Block and Fodor (1972) p.240.

supervenience principles that cannot be accounted for in either of the two ways we have discussed.

1.5 The Justification of Physicalism

Why should we believe physicalism? There are primarily two kinds of reasons that physicalists have provided. Both depend on the astonishing success of physics to date. The first reason is provided by the particular reductions of higher level predicates to lower level, and ultimately physical predicates that have either been carried out or which, it is claimed, it is reasonable to believe could be carried out. As we saw, such reductions ground and explain certain instances of E that are required by P. The most developed examples are the reductions of thermodynamic and chemical predicates and processes to quantum mechanics and the reductions of certain biological predicates and processes to chemical predicates and processes (e.g., biochemical accounts of photosynthesis, cell growth, genotype transmission). For macro-processes other than those involving mentality there is little reason to doubt that such reductions exist.

These reasons for physicalism are not overwhelmingly strong. Anti-physicalists might grant that non-mental macroscopic properties are micro-physically reducible but claim that mental properties are different; that there are in principle reasons why they are not reducible in the manner of non-mental macroscopic predicates to physical predicates. The conceivability

argument is supposed to establish just this. But I think it is fair to say that, pending a persuasive anti-reductionist argument, the successful reductions that have been produced provide reason to believe that the truth of statements involving mental predicates is not incompatible with physicalism.

The second argument for physicalism is stronger. It begins with the premiss that physics is causally closed.[32] To say that physics is causally closed is to say that every physical event has as a full causal explanation in terms of prior physical events and the laws of physics.[33] If F is a higher level property which has different physical effects depending on whether or not it is instantiated, e.g., a physical detector of the presence or absence of F in a region is physically possible, then it can be shown that if F is instantiated at some time t in R then the complete physical description of the world metaphysically entails that F is instantiated at t in region R. The argument is a reductio. Suppose that F's being instantiated at t fails to supervene on the

---

[32]The idea behind the argument is suggested by McGinn (1982, p. 29) and formulated in slightly different ways by Papineau (1993b, pp. 17-20) and Loewer (1995). An error in their line of reasoning is found by Witmer (1997, pp.195-304) who makes the needed repairs.

[33]By "physical event" I mean the values of fundamental physical quantities in a region of space at a time (I am ignoring relativistic complications). If the fundamental laws are indeterministic, then causal closure says that prior physical events and laws specify the probabilities of physical events and these probabilities remain the same in the face of conditionalization on statements not couched in the vocabulary of fundamental physics.

physical facts at t. Then whether or not F is present can make no causal difference to the subsequent physical events, since they are determined (up to objective indeterminacy) by the physical facts at t. Since it is plausible that every genuine property, including intentional and phenomenal properties, does make a causal difference to physical events, it is plausible that every property, and so every truth, supervenes on the complete physical description of the world.[34]

What would it be like if phenomenal properties failed to supervene on physical facts? These seem to be three possibilities:

> 1. P is false, but physics is causally closed. In this case phenomenal properties are epiphenomenal with respect to physical events. (Epiphenomenalism)
>
> 2. P is false and physics is not causally closed. Phenomenal properties have their own causal powers that are not derived from supervenience relations on physical properties. (Dualist interactionism).
>
> 3. P is true, but phenomenal properties are not instantiated. (Eliminativism)

None of these positions are attractive. Epiphenomenalism makes it puzzling how we can come to know whether the epiphenomenal property is exemplified. And, in any case, phenomenal properties, e.g., being in pain, certainly seem to have physical effects. Dualist interactionism has been

---

[34]The complete physical description, of course, have to includes the laws.

proposed to deal with the measurement problem in quantum theory and for other reasons as well. It is probably correct to say that we do not know for certain that it is false. On the other hand, there is no evidence whatsoever in favor of it, and there are no interactionist theories even sufficiently articulated to test. Eliminativism is difficult to take seriously. I will put off doing so to Chapter Two.

In any case, an argument would have to be very compelling to lead one to accept one of the above alternatives. The conceivability arguments claim to compel us to do just this. But before turning to them I want to look a little more closely at the nature of consciousness.

CHAPTER TWO:
CONSCIOUSNESS

2.1 KINDS OF CONSCIOUSNESS

Consciousness is a complicated phenomenon involving a number of different aspects. Ned Block (1994a, 1994b) provides a useful taxonomy. His main distinction is between what he calls “cognitive consciousness” and “phenomenal consciousness.” Cognitive consciousness involves a mental state’s possessing a particular kind of representational content or a particular cognitive role. Block distinguishes among various kinds of cognitive consciousness. A state is *access conscious* if its content is inferentially "promiscuous", in the sense that it freely enters into reasoning involving other propositional attitudes, and is available for the rational control of behavior (verbal behavior included). A mental state is *self-conscious* if it involves a first-person representation of one's self. It is *reflectively conscious* if one has a "higher order" thought about that state to the effect that one is in that state.[35] According to Block, when we say of some thought that it is conscious we might mean any of these aspects.

---

[35]Subpersonal mental representations posited by cognitive

Phenomenal consciousness, on the other hand, involves experiential quality, the “what it’s like” feature possessed by some mental states and processes. For example, when listening to, e.g., a Bartók string quartet, there are various auditory and other sensations, feelings of excitement, agitation and so forth that partly make up my experience. On any particular occasion, there is something it’s like to have these experiences. Phenomenal consciousness is a determinable - there being something it’s like -, with various kinds of determinates, i.e., the specific ways it is like. In the philosophical literature on consciousness these determinate kinds of phenomenal consciousness are called ‘qualia’.[36] There are qualia associated with the various senses, i.e., auditory qualia, visual qualia, etc., and also distinctive kinds of qualia associated with various emotions, reflective thinking, meditation and so forth.

A typical human conscious episode, say the pain a person experiences when she has a headache, involves both phenomenal and

---

psychologists in theorems of language processing (see, e.g., Chomsky 1975), vision (Marr 1982), etc., are unconscious in all of the above ways. Mental states posited by Freudian theory are not access conscious (access is available only with the help of your analyst) but may be self-conscious and reflectively conscious. It is not implausible that my cat’s mental states are access conscious but not self or reflectively conscious.

[36]‘Qualia’ is used to refer to both determinate phenomenal qualities, and to token instances of determinate phenomenal qualities. I will mostly use the term to refer to the property. Occasionally, I will also use ‘qualia’ in an adjectival sense, as in ‘qualia property’, to point to the determinable (i.e., phenomenal property).

cognitive consciousness. There is the phenomenal feeling of the headache: there is something it's like to have it. This feeling will typically involve different sensations and will change over time. The state is typically access conscious, since it may cause her to decide to take an aspirin, and typically self and reflectively conscious since she will judge that she is experiencing a headache.

For all I have said so far, cognitive and phenomenal consciousness may be two aspects of the same phenomena, or involve phenomena that are metaphysically or nomologically linked. However, Block argues that cognitive and phenomenal consciousness can be pried apart and are not just two aspects of the same phenomenon. He claims that it is possible for a mental state to be, say, access conscious, but not phenomenally conscious, and *vice versa*.

One of his arguments for the possibility of access without phenomenal consciousness involves blind sight. Certain people who are blind in that they experience no visual sensations, still apparently can obtain information about their environments via vision. In such cases Block thinks it is plausible that there is at least partial access to visual information, but there is reason to think that it is not accompanied by phenomenal consciousness since the subjects report that they experience no visual sensations.

The existence of phenomenal consciousness without cognitive consciousness is more speculative. One of the tentative examples Block (1994) gives of phenomenal consciousness unaccompanied by access consciousness has to do with reports of phenomenally conscious events under general anesthesia. Patients claim that the operation hurt. Another example comes from a study done on pilots during WW II by American dentists. The un-pressurized cabins caused the pilots to experience sensations in their teeth that could be interpreted as some kind of recreation of the pain of previous dental work. The pilots' recreation of previous pain would only follow dental work done in general anesthesia; procedures done in local anesthesia do not leave such "bodily memories".

While it seems that it is metaphysically possible (and actual if Block's interpretation of blind sight is correct) for there to be states that are cognitive conscious but not phenomenal conscious [37], the reverse is less obvious. The very concept of a phenomenal state that is not accessible and/or not available to second order judgements is quite peculiar. How can a state possess *what-it-is-like-ness* without it being like something for someone? In other words, phenomenal consciousness seems to be essentially subjective. And that seems to involve that it is available to the subject's cognitive

[37]Assuming, of course, contrary to some (e.g., Rosenthal 1990), that phenomenal consciousness cannot be reduced to cognitive consciousness.

system.[38] To think otherwise, i.e., that there are phenomenally conscious states that are not access conscious is to countenance cognitively isolated "islands" of phenomenal experience. There would be little more reason to say that they belong to a particular person rather than to her toes. There seems no reason to say that they are part of her mind.

Be that as it may, the important point is that it is phenomenal, and not cognitive consciousness that is relevant to the Conceivability and Explanatory Gap Arguments. The concept of cognitive consciousness is the concept of something that fulfills a certain role. Assuming that representational content poses no problem for physicalism - a big assumption -, there seems to be no special problem for physicalism posed by cognitive consciousness. The problem is the usual one of finding those physical, plausibly neurophysiological, structures and processes that implement the cognitive consciousness roles.

---

[38]Jennifer Church argues along these lines in an unpublished talk; Rutgers 1996.

But this would not solve the problems posed for physicalism by phenomenal consciousness. An account of cognitive consciousness does not, by itself, explain the nature of phenomenal consciousness or shed light on how it can be physically implemented. In any case, the Conceivability Arguments, if sound, would show that phenomenal and cognitive consciousness are metaphysically distinct, since a being with cognitive consciousness, but lacking phenomenal consciousness, certainly seems conceptually possible.[39]

2.2 PHENOMENAL CONSCIOUSNESS

Human beings, and no doubt other creatures, are capable of experiencing an enormous number and variety of qualia. Below is a very incomplete list to remind us of the enormous richness of phenomenal experience.

---

[39] The fact that it may seem (pace the argument I gave above) that it is conceptually possible for there to be phenomenally conscious states that are not cognitively conscious may encourage the thought that this is a genuine metaphysical possibility. This is an instance of the kind of conceivability argument that I will be discussing and undermining in subsequent chapters.

A) Bodily sensations: pains, aches, itches, tensions, tickles, tingles, pleasures.

B) Perceptual sensations: the smell of a rose, the sight of a cloudless sky, the sound of a middle C, the taste of a ripe apricot.

C) Images: dream images, fantasy images, memory images, after images.

D) Feelings: rage, longing, contentedness, boredom.

E) Moods: amusement, elation, anxiety.

I want to investigate now what we can say about qualia in the spirit of asking *how they seem* to us, while recognizing that we might be mistaken about some of the features that we are inclined to attribute to them. I have gleaned these from the philosophical literature, but they also mostly strike me as what a non-philosopher would say; if not quite in these words.

1) Determinateness of what it's like. I, and I am sure every other person, have the impression that qualia are determinate properties with respect to what it's like to instantiate them. For example, in experiencing a toothache there seems to be a determinate property of that toothache, and it also seems that I can experience toothaches of the same type at other times and that other people can experience toothaches of the same type. Further, it seems to me that other toothaches may be more or less similar to this one with respect to *what this toothache is like;* i.e., *in qualitative content.* Qualia seem to exhibit a complicated structure of similarities and differences in

qualitative content. Headaches are more like toothaches then visual experiences of red, while the latter are more like visual experiences of green than like headaches; tastes are more like each other than like visual experiences and the taste of a lemon is more similar to the taste of a grapefruit than to the taste of Hungarian sausage.

2) Subjectivity. Qualia are subjective properties. An experience is always an experience for someone; it is an experience from the subject's point of view.

3) Asymmetric epistemology. Each person has epistemic access to her own phenomenal states via introspection, but to other peoples phenomenal states only via behavioral or other physical intermediaries. Further, one can know what an experience is like only by having that experience.

4) Intrinsicness. Qualia appear to be intrinsic properties of mental states. It is a very strong intuition that a brain in a vat or a swampman can experience exactly the same qualia that I am experiencing now.

5) Categoricalness. Qualia appear to be categorical properties. By that I mean that they are properties that are not individuated in terms of their causal and nomological dispositions. A paradigm non-categorical property is **solubility**. Its instances are grouped together in virtue of a causal

disposition, i.e., to dissolve in water. Qualia do not seem to be like that. Instances of qualia are grouped together in virtue of similarity of quality.

6) Intentionality. Qualia, or at least many types of qualia, seem to be essentially intentional. The visual qualia I experience when looking at the sunset seem to have intentional content. They represent colored expanses in my visual field and via them they represent the setting of the sun. Further, it is not merely that these qualia contingently, or conventionally represent what they do in the way, for example, the word “red” represents **red**. The representational features seem to be essential to the qualia.

7) Collapse of appearance/reality distinction. For qualia, the usual distinction we make between the way something appears and the way it really is, collapses. For example, there is a distinction between it appearing to me that there is a cat at my feet, and there being a cat at my feet. Either may obtain without the other. But the situation seems different with qualia. If it appears to me that I have a headache then I have a headache. Of course, to avoid misunderstanding, I should say that I am using “headache” to characterize a kind of qualia intrinsically with no implication as to what its causes are. So understood, to appear to have a headache is to have a headache.

8) The conceivability of zombies and inverted spectra. No matter what physical information I have about a person, about her brain states and

processes, her functional structure and dispositions to behave, there seems to be no contradiction in her being like that and yet instantiating no qualia. Also, there seems to be no contradiction in there being two persons who are physical duplicates but who instantiate inverted qualia. In particular, where one person experiences phenomenal green, the other experiences phenomenal red and vice versa.

9) Qualitative character. One of the most obvious features of qualia is that there is something it's like to have them; they have a certain specific phenomenal feel.

Eliminativism

Reflection on the characteristics we are inclined to attribute to qualia leads some philosophers to think that no physical or functional property can possess them.[40] If one thinks that and is also committed to physicalism then

[40]The conceivability arguments are intended to establish that qualia are not physical or physically realizable. But even without argument, one may well wonder how there can be no distinction between appearance and reality for any physical property. In fact, the problem doesn't really have to do with *physics* but rather with *objectivity*. Any property for which no appearance/reality distinction can be made must be very different from familiar properties. Their subjectivity must be essential to them and it is hard to understand how any property can be like that. So this consideration does

one might endorse *Eliminativism*; the view there are no qualia or that qualia are not instantiated.

However, the proposal that there are no qualia is likely to strike all but the most sophisticated philosophers as absurd. The reason that eliminativism strikes one as absurd is that our acquaintance with qualia - our knowing what various experiences are like - is direct. Because of this we cannot explain it away as *mere appearance.* It *is* appearance. Contrast this with, for example, the apparent immobility of the earth and the revolution of the sun around the earth. We can explain why the earth appears to be still (though it moves) and why the sun appears to orbit the earth (though it is the other way around). Even thought it is ”obvious”, or at least it was to our ancestors, that the earth does not move there is logical distance between the thought that the earth does not move and the thought that it appears that it does not move. Because of this we can understand how it may really be that the earth moves even though it appears not to.

---

not favor dualism over physicalism; it rather supports the idea that qualia must be eliminated.

But in the case of qualia it makes no sense to say that I merely appear to have them but really don't. If I know anything, I know what my experiences are like. Further, qualia are absolutely central to one's life. if I thought that, as of midnight, my physical and functional organization would be similar to how it is now, but that I would no longer have qualia, I would regard that as death, or a kind of living death.[41] My body would go on living as a zombie, but I would be no more.

Eliminativism then is a non-starter. There are qualia. But accepting that there are qualia does not commit one either to qualia actually being everything they appear to be or to their being scientifically important properties. It may be that some of the ways that qualia appear, e.g., their appearing to be categorical is mere appearance. What cannot be mere appearance is that there is something it is like to have a headache, see a cloudless blue sky, etc.

[41]This thought may, of course, express a metaphysical impossibility.

Qualia are important properties from a subjective point of view. They are, as we mentioned above, absolutely important to our lives. But that is entirely compatible with their being not very important from a scientific point of view. On the abundant conception of properties, not every property will be scientifically important; occur in laws or scientific explanations. Some philosophers seem to think that by showing that the property has no role to play in science, in particular, is not invoked by any explanatory theory, one has thereby shown that the putative property does not exist (and so does not have instances). This may be a reasonable stance if one has a sparse (or elite) conception of properties.[42] But on that conception it is unreasonable to say that a thought of the form *There are Fs* is true only if *F* refers to an (elite) property. And you still have to give an account of what makes thoughts of the form *There are Fs* true.

If there are qualia, and if physicalism is true then true qualia thoughts are made true by physical facts. But how are they made true? A completely satisfactory way of answering this would be to specify a physical or functional property and then show for each condition on our list either that this property satisfies that condition, or, if it fails to satisfy the condition, explain that away as mere appearance. There are a number of attempted reductions in the

[42] Recall the discussion of properties in Chapter One.

literature identifying qualia with certain neurophysiological, computational, and representational properties. At best, some of these theories provide partial accounts . None provides a persuasive account of how the *what it's like* feature of qualia results from arrangements of properties that are not qualia. So, rather than examining these accounts of the nature of qualia, here I want to look at an account of the nature of qualia *concepts* due to Brian Loar (1990, 1997).[43]

Loar's suggestion is that many of the characteristics, especially the most problematic ones, that we are inclined to attribute to qualia are actually due to the way we think about our own qualia; to the nature of the concepts that we employ to pick out our own qualia.

2.3 The Nature of Phenomenal Concepts

[43]Sturgeon proposes ideas similar to Loar in Sturgeon (1994).

Phenomenal concepts are concepts that a person applies directly to her qualia. They are tokened when, for example, a person sips a red wine and notices first sensations of tanginess and then a sensation of sourness. Loar (1990, 1997) says that phenomenal concepts belong to a wide class of concepts he calls "recognitional concepts." Recognitional concepts are those that enable their possessors to perceptually recognize instances of the concept under certain circumstances.[44] Thus recognitional concepts are connected, via their inferential roles, with basic perceptual concepts, sensory inputs, images, etc. Loar says that recognitional concepts have the general structure "is of that kind" where the demonstrative purports to refer to a kind as exemplified through a perception or image of an instance of the concept.

Here is an illustration. Jerry sees a platypus for the first time in the zoo, and forms the concept *animal of that kind*, where the demonstrative is focused on its reference by his perception. His concept is connected with

[44]Fodor argues that there are no recognitional concepts, not even *red* (Fodor 1997, Ch. 4). Fodor's arguments against recognitional concepts depend on considerations of compositionality. He thinks the features that supposedly make a concept recognitional do not compose, and so cannot really be constitutive of the concept. But even if that is correct about concepts in general, I will argue that *phenomenal* concepts (like indexicals, demonstratives, and logical terms) are different, in that they are hard-wired to occupy a certain conceptual role, which underlies the recognitional capacities associated with them.

other concepts that constitute the beliefs he forms on the basis of the perception. For example, it is connected to the concepts *web-footed, aquatic, brownish, etc.* Possessing this concept, he is able to recognize other instances.

Loar mentions a number of features of typical recognitional concepts that are important to mention.

1) they are demonstrative and perspectival;

2) they enable a thinker to recognize their instances via perception;

3) the perceptual concepts via which they are applied are typically conceived of as contingently connected to the kind referred to by the concept.

Phenomenal concepts are a special kind of recognitional concept. Their basic application is to one's own phenomenal states as they occur, e.g., *itch again*. Of course, we also apply phenomenal concepts in memory and in reasoning and to other people. These applications are derivative, depending on the first person, present tense applications, so I will describe the basic application first.

Unlike other recognitional concepts, a phenomenal concept does not refer via a continent mode of presentation. Instead, it is applied *directly* to an internal state. Loar suggests that a phenomenal concept has a mode of presentation that is essential to its reference. What he seems to have in mind is that when tokening a phenomenal concept, the reference is focused

or determined by a token of that very kind of phenomenal state. The mode of presentation is then the very property referred to. For example, when tokening *pain*, the mode of presentation is the *painfulness* of the token of pain to which the concept is applied. Thus the mode of presentation is essential to the referent.[45]

There is a puzzle about the preceding account. If a phenomenal concept, e.g., *pain*, has reference only in virtue of the occurrence of a painful state, then it looks as though referenceless basic tokens of *pain*, or rather, of a concept just like it except for not having a reference, are possible. In that case, a person could token a concept *pain*$^{\#}$ that is just like *pain*, minus the phenomenal state. It could even appear to the person that she is in pain, and the tokening of *pain*$^{\#}$ might cause pain behavior, beliefs, memories, etc. It seems as thought on the account so far sketched, the relation between a phenomenal concept and its reference, though close, is not yet sufficiently intimate. As we mentioned earlier, in the case of qualia, the appearance/reality distinction collapses (this is feature 7 on our list).[46] If,

---

[45]The mode of presentation of a concept, as we mentioned in Chapter One, is how the thinker conceives of the concept's purported reference. It is a determinant or partial determinant of the concept's reference. For most concepts it can be characterized descriptively. But for phenomenal concepts, it is the reference.

[46]This is sometimes expressed by saying that our phenomenal judgements are *incorrigible*.

however, the tokening of the concept *pain* (or rather, the internally indistinguishable concept *pain*$^{\#}$) can come apart from instantiations of the property **pain**, then the appearance/reality distinction applies to phenomenal concepts just as much as to every other concept.

Here is a suggestion to overcome that. Suppose that a phenomenal concept of a given type is identical to its reference, e.g., the concept *pain* is identical to the property **pain.**. A particular token of the concept *pain* is a particular painful state. On this account phenomenal concepts are *self referring*; something of the sort *internal state of the phenomenal kind which I exemplify.*[47] This is, in a way, in agreement with Dennett when he declares:

> "You seem to think there's a difference between thinking (judging, deciding, being of the heartfelt opinion that) something seems pink to you and something *really seeming* pink to you. But there is no difference." (1991a, p. 364)[48]

---

[47]The pronoun "I" is intended to be understood as the concept referring to itself. The suggestion that phenomenal states are self-referring phenomenal concepts bears some similarity to Tyler Burge's (1988) account of self-knowledge. According to Burge, certain judgements about the intentional contents of one's states are self-certifying. Take for example, *I'm now judging that I am thinking that cats purr.* In order to make the judgment one has to do the thinking so the judgment must be true. On my proposal, in order to token the phenomenal concept (in the basic way), one has to token the phenomenal state to which it refers.

[48]Of course, his view is that what this means that there are no qualia, that qualia are not real. As he puts it, "[t]here is no such phenomenon as really seeming - over and above the phenomenon of judging in one way or another that something is the case." (ibid, p. 364) My view is that qualia are perfectly real, *and* that our qualia-judgements *are constituted*, at least partly, by our qualia-states.

Of course, a given phenomenal state will exemplify many phenomenal properties. The suggestion is that it refers to the most determinate qualia it exemplifies.[49] Note that the suggestion is not that all there is to phenomenal concepts is that they are self-referential. They have other intrinsic properties and conceptual roles. It is these other features that distinguish various phenomenal concepts (and phenomenal states, as these are phenomenal concepts) from each other. The preceding proposal is speculative and not part of Loar's account. But it seems to have the advantage that it can account for most of the apparent features of qualia, listed in 1-8 above.

What about non-basic applications of phenomenal concepts? Clearly, a person can token a concept that refers to pain without herself literally experiencing pain as when she replies to her dentist's question by "I am not in pain" or when one sees another person stub her toe and thinks *that must really hurt*. These concepts must be distinct from basic phenomenal concepts. Here is a suggestion concerning how they work. A person forms a memory of a basic tokening of, e.g., *pain*. This memory contains a concept that refers to the basic tokening, i.e., to the particular pain. The derivative concept *pain* has the content *same phenomenal property as that* where the demonstrative demonstrates the memory of pain.

---

[49]Or perhaps it refers ambiguously to each of the qualia it exemplifies.

The preceding account of phenomenal concepts has some interesting consequences. I will list these below, matching them with our list of properties that qualia appear to possess.

1) Determinateness of what it's like. Phenomenal concepts present their references as determinate properties. Instances of, e.g., *appears blue*, all seem to have something phenomenally in common, and seem more similar to instances of *appears turquoise* than *appears red.*

2) Subjectivity. Phenomenal concepts present their references from a subjective point of view. The basic concepts can apply only to a state of the subject, and derivative phenomenal concepts make reference to the basic applications.

3) Asymmetric epistemology. The nature of phenomenal concepts accounts of unique epistemic access each person has to her own qualia.

4,5) Intrinsicness, categoricalness. Basic phenomenal concepts present their references as intrinsic and categorical. A phenomenal concept is plausibly a relational and dispositional property since it is constituted in part by its causal relations to other concepts. Even so, it presents its reference, i.e., itself, without disclosing these relations or dispositions. And it gives the impression that its instances are grouped together in virtue of qualitative similarity (since they are instances of the qualia), not causal dispositions.

6) Intentionality. Qualia are intentional since they are concepts that self-refer. This, in a way, captures our intuition that qualia are representational, although it is important to point out that the fact that qualia concepts are representational in the way I am suggesting, is *not* a priori available to us. Further, the fact that they represent themselves (or rather the phenomenal property they exemplify) does not preclude them from representing other properties as well. Thus, a visual sensation of blue may represent the qualia blue, as well as the blueness of the sky. The account also makes for a unified theory of mind on which all mental states are intentional. One problem with theories on which qualia are not intentional states is understanding how they can be part of the mind any more than other bodily properties.

7) Collapse of appearance/reality distinction. It explains the collapse of the appearance/reality distinction. Since qualia are self-referential there can be no distinction between, e.g., a state appearing to be painful (when tokening the basic concept *pain*) and its really being painful! Finally, we can see why we find eliminativism - the idea that phenomenal concepts have no reference - so absurd. If phenomenal concepts are self-referential and we have even an implicit appreciation of that then the proposal that they lack reference will be defeated merely by tokening a phenomenal concept.

But this does not mean that phenomenal concepts denote scientific properties. It may be that the various tokens of a phenomenal concept, e.g., the concept *tickle* have very little in common from a neurophysiological point of view. While tickles are grouped together in a particular way on the basis of our phenomenological judgements of similarity, the states judged similar phenomenally may be a physical hodgepodge. So the view I have sketched is compatible with eliminativism if that doctrine is understood as claiming that phenomenal properties will find no special place and no important place in scientific theory. But who would have thought otherwise? And their not being important from the point of view of science hardly means that they are not important at all even to the scientific enterprise.

8) The conceivability of zombies and inverted spectra. It explains why zombies and inverted spectra are conceivable. The reason lies in the fact that basic phenomenal concepts are direct recognitional concepts (and derivative phenomenal concepts make reference to their direct cousins). Because they are direct recognitional concepts a subject applies them directly to her internal states and not in virtue of those states satisfying certain other features, e.g., satisfying some functional role or being related to physical phenomena in such and such a way. Because of this, there are no constitutive conceptual connections between phenomenal concepts and behavioral, functional, etc. concepts that are sufficient to yield a conceptual

entailment from a physical-functional description to a phenomenal description. Further, on this account it may be that distinct qualia, e.g., *appears red* and *appears green* possess perfectly symmetric conceptual roles. If so then qualia inversion will be conceptually possible.

So far as I can see, Loar's original account of phenomenal concepts also explains why we are inclined to attribute the conditions on our list to qualia; except 6 and 7, intentionality and incorrigibility, which my version helps explain better. There is one item on the list however, which neither account explains. This one is the most significant: the qualitative feel of qualia, e.g., that there is something it's like to see a cloudless blue sky (9 on our list). But, as we will see, the accounts explain why we cannot provide a satisfactory reduction of qualitative feel. I return to this in the conclusion.

Loar's account of phenomenal concepts is neutral concerning whether or not qualia are physically realized properties. They are, if the account is correct, at least partly functional properties since they are concepts and concepts are partly functional properties. But they may have other intrinsic features incompatible with physicalism. The account is, we might say, a double edged sword. On the one hand it accounts for many of the features of qualia in terms of *the way we conceive* of them. This is quite favorable for physicalism. On the other hand, it provides an account of phenomenal concepts that implies that zombies are conceptually possible. This is the lead

premise of the conceivability arguments. If these arguments are sound then phenomenal concepts do not refer to physically realizable properties. So we now turn to examine the conceivability arguments to determine whether or not this double edged sword is capable of cutting the “World-knot” as Schopenhauer termed the Mind-Body Problem.

CHAPTER THREE:
THE CONCEIVABILITY ARGUMENTS

In this chapter I will be outlining and explicating expositions of the Conceivability Argument, as well as a close cousin of them, the Gap Argument. I will be pointing out the common assumption that runs through these arguments; this will prepare the ground for the Master Argument which will be adaptable to refute all of the Conceivability Arguments. I will be critical in my exposition; however, I will not present a refutation of these arguments at this point. The refutation of all the Conceivability Arguments will be the subject of Chapter Four.

3.1. DESCARTES' ARGUMENT FOR THE REAL DISTINCTION BETWEEN MIND AND BODY

Descartes has been the originator of a long line of arguments called the Conceivability Arguments for Dualism. Descartes famously argued (*Sixth Meditation* in: Cottingham 1984, II. p. 54) that since it is possible to conceive of his mind and body existing separately they are really distinct. Descartes made the connection between conceivability and possibility *via* the notion of essence.

Certain philosophers, among them Descartes, hold an *essentialist* view. They hold that things, events, and properties have a nature that they

necessarily have: it is *metaphysically necessary*, for example, for material bodies to be extended, for minds to be thoughtful, or, as modern essentialists hold, for water to be composed of two hydrogen and one oxygen atom, as it is of their nature to be so. The notion of a nature of things, events, properties, etc., has played an important role in the philosophical investigation of what there is, how entities are constituted; recently it plays a role in understanding and explicating theoretical reduction, and many other subjects. Conceptual possibility, and necessity, on the other hand, concerns what is coherently *conceivable*, not as a matter of psychological fact, but rather as a matter of our concepts.

Descartes made the connection between conceivability and necessity/possibility in the following way. He thought that when we conceive of, e.g., a substance, clearly and distinctly, we have a *complete conception* of that substance. Having a complete conception of a substance means that we conceive of it *through its essence*, and, importantly, *through its whole essence*. He thought that when we conceive of bodies clearly and distinctly, we see, by our very conception of bodies, that their whole essence is to be extended, and when we conceive of minds clearly and distinctly, we see that their whole essence is to be thoughtful. Descartes then argued that we can conceive of bodies clearly and distinctly without the property of thought, and, reversely, we can clearly and distinctly conceive of minds without extension.

This, however, means that it is possible for minds to be disembodied, and reversely, for bodies to exist without minds, since it is of the essence of minds to be thoughtful, but it is not of the essence of bodies to be so; and, it is of the essence of bodies to be extended, but not of the essence of minds to be so.[50]Minds and bodies are really distinct. The argument, reconstructed from the *Sixth Meditation*, goes like this:[51]

1) Whatever I can clearly and distinctly understand can be brought about by God.

2) If α belongs to the essence of A and β belongs to the essence of B, and I can clearly and distinctly understand B to exist without α and A to exist without β, then I can clearly and distinctly understand A to exist without B and B to exist without A.

3) Thought belongs to the essence of my mind, extension belongs to the essence of body.

4) I can clearly and distinctly understand body to exist without thought and my mind to exist without extension.

5) By (1), (2), (3) and (4), my mind can exist apart from body.

6) If A can exist apart from B, and vice versa, A is really distinct from B.

7) Hence, by (5) and (6), mind is really distinct from body.

---

[50]Actually, strictly speaking, he only argues that *if he has a body*, then it is separate from his mind. A plausible way to read his argument is that he was trying to prove that he (or his mind) could exist in the absence of anything else existing.

[51]I am using Margaret Wilson's interpretation of the argument (see Wilson 1978).

This is a classic statement of Dualism.[52] First, notice that questions about the existence of God are important for Descartes' overall position; however, from the point of view of this argument these questions are irrelevant. The argument could be stated without invoking claims about the power of God: the possibility of God's bringing about something can be thought about in terms of possibility *simpliciter*. Premiss 1 in this interpretation would then simply state that conceivability implies possibility.

The most controversial premiss, and the one that has gotten the most attention is premiss 1. Many philosophers deny that conceivability implies possibility. Descartes was criticized early on by Arnauld for assuming that the essential properties of his mind and body do not go beyond what he is aware of:

> How does it follow, from the fact that he is aware of nothing else belonging to his essence, that nothing else does in fact belong to it? (in John Cottingham; Robert Stoothoff; Dugald Murdoch 1984, Vol II, p. 140)

Descartes did assume that clear and distinct perception delivers the sole essence of substances. He held that our clear and distinct perception of

---

[52]Descartes' argument appeals to mentality as a homogeneous phenomenon. In the more current literature the argument is usually formulated with respect to a particular aspect of the mind: i.e., the experiential, qualitative aspect. Other aspects, like intentionality, seem to pose less of an insurmountable problem for naturalization.

substances is, in a certain sense, *transparent*. Let's call it the *Cartesian Transparency Thesis* (*CTT*):

> (*CTT*) When you clearly and distinctly perceive of substance A and substance B, that is, you conceive of them *through their whole essence*, and still do not see that they are *the same substance*, then they must be different.

This assumption, however, is contentious; Arnauld's objection has to be addressed. In different forms, but this very same objection applies to all the Conceivability Arguments, and the inability to answer satisfactorily will be their ultimate undoing. However, I defer detailed discussion until later.

There is another argument in Descartes that we might call the Argument From Doubt. In the *Second Meditation* he suggests that Dualism is true since he can suppose that his body does not exist, at the same time that he is certain his mind exists. The argument is the following.

1) I am certain that my mind exists.

2) I am not certain that my body exists.

3) My mind is diverse from my body.[53]

---

[53]Descartes actually runs the argument a little differently. He starts by observing that he is certain he *himself* exists, whereas not certain that his body exists. However, since he thought he knew that he was essentially a thinking thing, a *mind*, my formulation, I hope, preserves the spirit of his argument.

This argument, as presented, is obviously invalid. A further premiss, like, e.g., that

2') If I am certain of the existence of *a,* but not certain of the existence of *b,* then *a* cannot be identical to *b.*

is needed to derive 3. However, this premiss is hard to defend. Nagel (1974) and Jackson (1982) developed similar arguments recently[54]; the problems they face are also similar. I will deal with these problems in detail in the following chapters.

Much of the contemporary discussion surrounding the metaphysics of mind originates in Descartes' work. He put the conceptual and epistemological rift between the mental and the physical center stage in his philosophy; and he was the first to draw clear ontological conclusions from it.

[54]The similarity is in the basic structure of the argument. All of these arguments move from the assumption that a certain type of knowledge-claim does not imply another to the conclusion that the subject-matter of the knowledge-claims is different.

3.2 NAGEL'S BAT ARGUMENT

In an article that is perhaps more famous than well-understood, Nagel (1974) presents a number of considerations that are meant to lead to the conclusion that we cannot see how phenomenal consciousness, or the subjective character of experience, could be reducible to any objective, and *a fortiori,* any physical feature of the world.

It is not completely clear what Nagel means by 'reducible' here. Reductive accounts, like identity theory, or functionalism, can be, as we have seen earlier, perspicuous, or non-perspicuous, merely stating a metaphysical relation. If Nagel's claim is that there is no *perspicuous* reduction of qualia, he is right, but this does not strictly imply Dualism. On the other hand, if he wants to argue, as I think he does, that there is no reduction, perspicuous or not, of qualia, his argument is not strong enough to support that claim. More on this as we proceed.

The title question, "What is it like to be a bat?"[55], urges us to reflect on the fact that in some sense of the word, we will never be able to know what certain experiences are like.[56] The experiences he cites in support of this

[55]I am going to call this paper, informally, the BAT. References to this paper will be to the version reprinted in Nagel (1979).

[56]One of the differences between Descartes' argument and the BAT is

claim belong to creatures quite alien to us, *viz.,* bats, but this seems to be incidental to his argument. The point seems to be that the *type* of experience in question is in some way inaccessible to us. Jackson's Mary, for example, as we will see in the next section, is in the same situation with regard to the experience of seeing red as we all seem to be with regard to, say, the bat's experiences associated with echolocation. And this is to say that we could be in this situation *vis a vis any* experience, possibly familiar or alien.

This reflection on inaccessible experience eventually is supposed to convince us that phenomenal properties are, at least seemingly, irreducible to physical properties. The reasons that he gives for this claim seem to have to do with the radical conceptual breach between phenomenal concepts and physical concepts. Nagel puts the point by insisting that we cannot *imagine* what it's like to be a bat. But, as Jackson argued, one cannot get an argument against Physicalism based on a claim about our imaginative powers (Jackson 1982, p. 131). The point is not about our *imaginative faculties,* it is about our *conceptual endowment.* Nagel's point really is that we do not (and cannot) have the right concepts that are needed to describe that bats' experiences, and so that we are not able to know what bats' experiences are like.

---

that while Descartes was concerned about *substances*, Nagel is concerned about *properties*, i.e., qualia properties and physical properties.

It appears to me that there are two strands of argument that run through the paper. Both arguments depend on a distinction between what Nagel calls 'subjective', and 'objective'. This is a distinction that keeps cropping up in all of Nagel's later writings as well (Nagel 1986, 1993).

Nagel spells out the distinction, as Colin McGinn (1987, p. 264) emphasizes, in at least two different ways. At one place, Nagel says:

> "It is beliefs and attitudes that are objective in the primary sense. Only derivatively do we call objective the truths that can be arrived at in this way." (Nagel 1986, p. 4)[57]

And then he proceeds to spell out what the objectivity of belief consists in:

> "To acquire a more objective understanding of some aspect of life or the world, we step back from our initial view of it and form a new conception which has that view and its relation to the world as its object...The old view then comes to be regarded as an appearance, more subjective than the new view, and correctable or confirmable by reference to it. The process can be repeated, yielding a still more objective conception."[58] (p. 4)

In other words, as Colin McGinn puts it,

---

[57]Although he is more explicit on this in Nagel (1986), the same distinction applies to the BAT-paper as well.

[58]There are also places where Nagel seems to imply that the subjective\objective distinction is really a distinction between appearance and reality. However, it is never spelled out how this distinction would figure in an anti-reductionist argument. It seems that there is no interpretation of the appearance-reality distinction that is helpful in this respect. We can either take this as a distinction between sense-data and physical *objects,* or we can take this as a distinction between veridical and non-veridical perceptual or belief *states;* but I cannot see in either case a ready tool to further the anti-reductionist cause.

$S_C\backslash O_C$ "a conception is subjective if it represents a fact *from* a specific point of view, exploiting that point of view as a medium of representation; otherwise it is objective". (McGinn 1987, p. 264)

On the other hand, it might be said that

$S_M\backslash O_M$ a fact or property is subjective if "it is part of (or essentially involves) a specific point of view", otherwise it is objective.[59] (McGinn 1987, p. 264).

We can call the two notions of subjectivity\objectivity $S\backslash O_M$ and $S\backslash O_C$, respectively (M for metaphysical, C for conceptual).

Let's now see the argument. Nagel asks us to try to contemplate the inner life of a bat. He observes that

1*) "...these experiences...[the sonar experiences of bats]... have... a specific subjective character, which is beyond our ability to conceive." (Nagel 1979, p. 170)

Nagel is going to draw a conclusion about what kinds of facts exist on the basis of this premiss. He draws the moral in the following way:

2*) "Reflection on what it's like to be a bat seems to lead us... to the conclusion that there are facts that do not consist in the truth of propositions expressible in a human language." (p. 171)

---

[59]In fact, the subjective-objective distinction spelled out in terms of *points of view* might not capture the sense of the distinction in which it has to do with *experience* and *qualia.* One might argue that there could be creatures that have perceptual concepts and beliefs but that have no experiences at all. Such creatures could have S_C concepts and S_M states, without having anything that we might want to call subjective experience. However, in the following I will ignore this point, since it will not effect my main criticism of Nagel.

2*) has very important consequences. Since there are, according to this line of thought, facts that the human conceptual repertoire is not rich enough to express, there are facts that go beyond basic physical facts (since we seem to possess no analogous handicap *vis a vis* physical facts).[60] The conclusion seems to be that there are more facts than physical facts. Reflection on the existence of experiences that we cannot conceive of, seems to lead Nagel to the denial Physicalism.

Let's try to reconstruct the argument a little more precisely.

1) There are no physical facts that do not consist in the truth of propositions expressible in a human language.

[60]Nagel is talking about the reducibility of properties and facts, whereas I introduced the notion of reducibility *vis a vis* concepts and thoughts (or terms and statements). I think talking about property reduction is misleading, since it commits us to saying that the property **water** reduces to the property **H_2O**, i.e., to itself. Better to say that the concept *water* reduces to the concept *H_2O.* I will ignore this difficulty.

2) There are experiences that we cannot conceive of.

3) Having an experience is a fact.

4) If there are facts that we cannot conceive of then there are facts that do not consist in the truth of propositions expressible[61] in a human language.

Lemma (by 2, 3, and 4):

5) There are facts that do not consist in the truth of propositions expressible in a human language.

So, by 1and 5,

6) There are non-physical facts.

Does this argument work? Nagel himself seems to be hesitant to draw this strong conclusion. In a later passage in the same article he puts his findings in the form of a cautious conditional:

> "...if the facts of experience - facts about what it's like *for* the experiencing organism - are accessible only from one point of view, then it is a mystery how the true character of experiences could be revealed in the physical operation of that organism. The latter is a domain of objective facts *par excellence* - the kind that can be observed and understood from many points of view and by individuals with differing perceptual systems." (Nagel (1979) p. 172)

In other words, his final conclusion is not that experience is *irreducible,* but that it is unfathomable *how* it could be reducible. It means that he does not whole-heartedly embrace his conclusion. What explains this vacillation?

---

[61]The expressibility Nagel has in mind here is not relative to human capabilities. A proposition that no human is capable to express might still be a proposition that is in principle *expressible* in a human language.

The answer seems to lie in the ambiguity of 1.

1) There are experiences that we cannot conceive of.

If "conceive" an experience here means "have an S_C concept" of the experience, then 1 is rather uncontroversial. We definitely do not have a conception that represents the bat's sonar experience "from a specific point of view, exploiting that point of view as a medium of representation" (McGinn 1987, p. 264). *A fortiori* we do not have the appropriate phenomenal concepts; which is a statement that few would contest. This construal of 1 will now be called 1'.

1') There are experiences we cannot have an S_C concept of.

1', however, provides no support for 3. For 3 to be plausible, an "ability to conceive" certain experiences would have to simply mean an "ability to entertain (any) concepts that refer to" those experiences. This second construal will be called 1".

1") There are experiences we cannot conceive of through any concepts.

The problem is, (1") is not so uncontroversial any more. (1") only follows from the uncontroversial (1') under the assumption that the only possible way to conceive of an experience (an S_M fact) is via a phenomenal (S_C) concept[62].

[62]Brian Loar (1990), takes the same line I am presenting here.

But no reason has been presented that an S_M fact can only be conceived through an S_C concept. After all, we often conceive of the very same thing via very different concepts; think of heat and molecular motion. So the strong conclusion follows only if the $S_C \backslash S_M$ distinction is conflated[63].

Nagel's hesitation to endorse the anti-reductionist conclusion is probably a reflection of his awareness of the $S_C \backslash S_M$ distinction, and an awareness of the fact that 1, on the innocuous reading, does not imply 3[64]. His conclusion is, as we have seen, not that subjective experiences are irreducible, but that we have no idea how they *could* be reducible.

It is instructive to see what he says elsewhere of this problem. In Nagel (1979) he puts the point like this: the subjective character of experience is not captured by any reductive account of the mental

---

[63]Another way to put the problem is that Nagel's argument only goes through on a transparent reading of 'conceives that'. However, the premisses are only plausible on an opaque reading. The Knowledge Argument has a very similar problem with respect to the ambiguity of our concept of knowledge. More about it in the next Section.

[64]At least it is so in the earlier works, like the BAT. In the later works, like Nagel (1986) and Nagel (1993), he abandons his caution and comes to endorse anti-reductionism. It appears that his arguments in these works are tainted by a failure to appreciate the very same distinction an appreciation of which originally kept him from coming down on the side of the stronger conclusion: the $S_C \backslash S_M$ distinction. I will not be able to argue for this in detail here.

> "for all of them are logically compatible with its absence. It is not analyzable in terms of any explanatory system of functional states, or intentional states, since these could be ascribed to robots or automata that behaved like people though they experienced nothing... The reason is that every subjective phenomenon is essentially connected with a single point of view, and it seems inevitable that an objective physical theory will abandon that point of view." (Nagel (1979) p. 166)

What Nagel appears to be saying in this passage is that because of a radical lack of connection between our subjective conception of experiences and our objective conception of physical or functional states, it will always seem possible that a physical (or functional state) occurs without the putatively reduced experience occurring. This means that the failure of any proposed reductive account will always seem possible. And so, even if the failure of the proposed account is *not* possible, we will never be able to see why. This, however, is a different argument from the one we have been examining so far: it sounds more like Descartes' Conceivability Argument, except that he does not quite draw the dualist conclusion. Here he argues from the apparent conceivability of absent qualia to their apparent possibility; whereas in the BAT he argued from his inability to have an adequate conception of certain experiences to those experiences being irreducible.

However, it would be a mistake to think that by the end of the BAT-paper Nagel gave up on the first line of argument and came down clearly on the side of some watered down Conceivability Argument *à la* Descartes. In the main part of the BAT-paper Nagel does not argue, in Cartesian fashion,

that in the case of *any* experience there is reason to think that it is non-physical, given that our physical conceptions of things seem to bear no relation to our phenomenal conception of things. He rather argues that, as long as there are experiences that are inaccessible to us, there is reason to think that there are non-physical facts. The argument, of course, depends on the non sequitur that we since we do not have a phenomenal conception of those experiences we do not have *any* conception of them. We are told that physical terms cannot refer to experiences that we do not have a phenomenal conception of.

If all he had in mind was a Conceivability Argument *à la* Descartes, then ordinary Dualism would have to seem to him a possible solution to the mind-body problem. However, Nagel makes it clear that in his view no Dualism that does not invoke the $S_M \backslash O_M$ distinction could provide a solution to the mind-body problem. He addresses this point with regard to mental substances:

> "The broader issue between personal and impersonal, or subjective and objective, arises also for a dualist theory of mind. The question of how one can include in the objective world a mental substance having subjective properties is as acute as the question how a physical substance can have subjective properties." ("Subjective and Objective", in Nagel 1979, p. 201)

This view, sometimes called "transcendentalism", only makes sense if what generates the mind-body problem, is something like the first line of

argument. If your problem is that no matter how many O_C concepts you have you still cannot seem to be able to conceive of the bat's sonar experience, then you will not be content to be told that there is a further O_M substance beyond the physical, namely, a mental substance that encompasses all experiences, the bat's experiences among them. Your problem was that the bat's sonar experience did not seem to be an O_M fact. If it was an O_M fact then you should be able to conceive of it, but you cannot and so anti-reductionism follows.

I do not endorse this argument but since it has certain close similarities with Jackson's Mary Argument, I will put off a detailed account of why it fails until the next section. Still, there is an important moral for the physicalist from this discussion: sooner or later one has to address the question of why the very idea of scientific reduction seems so problematic in the case of subjective experience.

3.3 JACKSON'S KNOWLEDGE ARGUMENT

Frank Jackson (1982) gives an argument that some think was anticipated obliquely by Nagel in the BAT-paper. I will try to show in what sense this is true. Jackson himself reflects on the connection between the two arguments:

> "Nagel speaks as if the problem he is raising is one of extrapolating from knowledge of one experience to another, of imagining what an unfamiliar experience would be like on the basis of familiar ones. In terms of Hume's example, from knowledge of some shades of blue we can work out what it would be like to see other shades of blue. Nagel argues that the trouble with bats et al. is that they are too unlike us. It is hard to see an objection to Physicalism here. Physicalism makes no special claims about the imaginative or extrapolative powers of human beings, and it is hard to see why it need do so.
>
> Anyway, our Knowledge argument makes no assumptions on this point. If Physicalism were true, enough physical information about Fred would obviate any need to extrapolate or to perform special feats of imagination or understanding in order to know all about his special color experience. *The information would already be in our possession.* But clearly it isn't." (Jackson 1982, p. 131)

Jackson emphasizes that the problem with qualia is not that we cannot *imagine* unfamiliar ones (as he sees Nagel as claiming), but that there is something we cannot *know* about them, unless we actually experience them[65]. His argument proceeds through a simple thought-experiment. He ask us to imagine Mary, a brilliant neuro-scientist who is forced to live in a black and white room. She maintains contact with the

---

[65]As we have seen in the previous section, a more charitable interpretation of what Nagel was arguing is that he maid a claim about conceptions, rather than imaginative powers. But we need not get into this here again.

outside world and conducts her research via a black and white TV monitor[66]. She specializes in the neurophysiology of vision and acquires *all* the physical information there is to obtain about color vision. Now Jackson ask us to consider what will happen when Mary is released from her black and white room. In particular, he ask us to consider whether she will *learn* anything about color vision. Jackson's answer is that she obviously learns new *facts.* But if this is so, then her previous knowledge was incomplete. However, she has *all* the knowledge there is about *physical* facts. Consequently, there are facts that are non-physical.

Here is the argument:

1) Mary knows all the physical facts.

2) Mary does not know what it's like to experience red.

3) What it is like to experience red is a fact.

Lemma:

4) There is a fact that Mary does not know.

Conclusion:

5) There are non-physical facts.

---

[66]Of course, there are some further details that have to be filled out for the story to work. E.g., we have to also imagine that she has been raised this way, and that her entire environment (including her body surfaces) is entirely colorless.

Now let's see in what way Nagel was anticipating Jackson's Knowledge Argument in the BAT Argument. Where Jackson assumes that there is no physical fact Mary does not know (premiss 1), Nagel assumes that there are no physical facts that cannot be conceptualized in a human language (premiss 1). Premiss 2 in Nagel's argument plays the same role as Jackson's premiss 2. Of course, there is a difference. Jackson's claim is that Mary does not *know* some facts, and Nagel claims that we cannot *conceive* of some facts. The function of these premisses, however, is the same within the overall structure of the argument; also, both notions lend themselves to ambiguities between opaque and transparent readings which will eventually undo the arguments. Finally, where Jackson claims that there is a fact Mary actually does not know (premiss 4), Nagel is claiming that there is a fact that cannot be conceptualized in a human language (premiss 5). The conclusion is in both cases that there are more facts than physical facts.[67]

Notice also that, understood right, Jackson's Mary Argument (like the BAT Argument) is a transcendentalist argument: if it works, it works against not just physicalism, but naturalism in general, i.e., against the view that all that exists is accessible from an objective, third person point of view. Let's suppose that what it's like to see red is a non-physical, irreducibly mental fact

[67]The difference, as we noted earlier, is that Nagel actually stops short of embracing the conclusion of his argument.

that can be described in some suitable, objective, mentalistic language. Let's suppose further that Mary in the black-and white room learns all the physical *and* non-physical, irreducibly mental facts (perhaps she reads a comprehensive encyclopedia of mentalistic facts and psychology). Mary still would not know what it's like to see red, having never left the black-and-white room.

The problem with Jackson's argument should be familiar by now: just like Nagel's argument only went through on a transparent reading of what it is to *have a conception* of a fact,[68] it only goes through on a transparent reading of what it is to *know* a fact. So it seems that Jackson has a further, implicit premiss:

1*) 'Knowing that' is a relation between a person and a fact.

[68]On a transparent reading of what it is to have a conception of a fact, substitutivity of identicals holds. More precisely, if F and G are the same facts then if one has a conception of the fact that F, then *eo ipso* one has a conception of the fact that G. The opaque reading this conditional does not hold.

Defenders of Physicalism locate the problem with this argument in different places. David Lewis (1990) and Laurence Nemirow (1990), e.g., objects to 3.[69] According to them, what it's like to experience red is not a fact. When someone comes to know what it's like to experience red, the knowledge acquired is not of the 'knowledge that' type, but the 'knowledge how' type.

Bigelow and Pargetter (1990), Dennett (1991), Loar (1990), Neander (1992), Papineau (1993) and others object to 1*. According to them, 'knowledge that' is a relation between a person, a fact, and a mode of presentation. This certainly seems right. Lois Lane, e.g., can know (i.e., before she knew that Clark Kent is Superman) that Superman is brave without knowing that Clark Kent is brave. The natural thing to say is that there is just one fact there, so a difference in facts cannot account for the difference in the knowledge claims. There is just one fact there and Lois knows it under one mode of presentation (*Superman is brave*[70]) but not under another (*Clark Kent is brave*). Notice that different modes of

---

[69]Actually, this answer to the Knowledge Argument is not entirely new. Herbert Feigl (1967) gave a version of it, together with a version of the Knowledge Argument. He drew a distinction between *acquaintance* with experience, which is not propositional, and *knowledge* by acquaintance, which is.

[70]I am using '*' to form names of modes of presentations from sentences.

presentation may engage action (and other thoughts) differently even though they are *of* the same fact.

Jackson can reply by granting that 1* is false and replacing it with

1**) 'Knowing that' is a relation between a person, a fact, and a mode of presentation.

But now he has to argue that the mode of presentation that Mary acquires when she learns what it's like to experience red cannot be *of* a physical fact. So the argument comes down to showing, that, e.g., *This is an experience of red* and *This is R-fibre firing* (or *This is an instance of functional property F*) are modes which do not pick out the same fact. But so far we have not been given an argument for this.[71]

Robinson (1993) takes a different tack: instead of showing that the two modes of presentations cannot be of the same fact, he tries to extend the original considerations to the modes of presentations themselves. He argues that the phenomenal mode of presentation of pain, according to the physicalist, would have to be itself physical. But, since Mary knows all the physical facts, she should know all the facts about the phenomenal mode of

[71]There is a very different kind of criticism of the Knowledge Argument made by, e.g., Watkins (1989). Watkins points out that Jackson's dualist conclusion, as Jackson himself admits, implies epiphenomenalism. However, if phenomenal properties, like experiencing seeing red, are epiphenomenal, how are we ever to *know* about them? So the Knowledge Argument seems to undermine itself. This criticism, as we have seen, applies to all the other conceivability arguments.

presentation, "including the fact of what that mode of access is phenomenally like" (p. 166).

However, this argument falls prey to the same objection that it tries to answer. Even assuming that the phenomenal mode of presentation of pain is a brain state type, possessing this brain state (and so knowing what it's like phenomenally to have that mode) is not the same as knowing that it is such and such a brain state. And the latter is all that is required to satisfy the claim that Mary knows all the physical (and neurophysiological) facts under their scientific guises. So the same objection that applied to the original argument applies to Robinson's version as well: we can know facts about a mode of presentation under different modes of presentation, and knowledge under the one does not imply knowledge under the other. This argument (as well as the Mary Argument) is inconclusive: nothing it says rules out a priori the possibility that an experience, *or* the phenomenal mode of presentation of that experience is physical.

There are a number of arguments that try to do the job so far undone by Mary Argument. If successful, they would show why a phenomenal concept and a physical concept cannot refer to the same property. They are i) the Property Dualism Argument, ii) Kripke's argument for dualism, iii) the Chalmers' and Jackson' new arguments. So these are the arguments to which we now turn.

Before we get to these arguments, let's see a very simple, but obviously unsatisfactory argument.

1) If P and Q are modes associated with the same state of affairs then it is knowable a priori that they are associated with the same state of affairs.

2) It is not knowable a priori that *This is an experience of red* and *This is P* are associated with the same state of affairs.

3) *This is an experience of red* and *This is P* are associated with different states of affairs.

Obviously, this argument is no good. *This is water* and *This is H_2O*, e.g., are associated with the same states of affair, but that is not knowable a priori, so 1 is false. To know a priori that they are associated with the same state of affairs, one would have to know a priori that water is H_2O under the mode *water is H_2O*. But that is not knowable a priori, even though water *is* H_2O. So, to have a convincing argument, we would have to be able to distinguish between those identity statements, like 'water is H_2O', which can be true in spite of the fact that they are not knowable a priori, and those ones (presumably 'what it’s like to experience red is P' is among them) which can only be true if they are knowable a priori. This is what the following arguments try to accomplish.

3.4 THE PROPERTY DUALISM ARGUMENT

The property dualism argument hangs on a particular view of modes of presentation. The semantics that is driving the arguments is Fregean in origin. On the Fregean view, meaning has two components, sense and reference. Associated with each term there is a descriptive sense graspable by each competent user of the term.[72] Sense is transparent, that is, it is knowable a priori whether two tokens are tokens of the same concept (where concepts are individuated by sense). Sense also determines reference, or, equivalently, sense determines an *intension* function f:W→R from possible worlds to referents. In other words, Frege thought that there are things that play both the role of determining reference and the role of distinguishing concepts. On his view, no two concepts can have the same reference in all possible worlds.

[72]Fregean sense is the prototype of what we have called ‘mode of presentation’; it is a theory of what it is that distinguishes different conceptions of the same proposition.

Following Frege, White (1986)[73] assumes that modes of presentation have two roles to play simultaneously. On the one hand, (i) they determine reference. On the other hand, (ii) they individuate concepts. He also assumes that modes of presentations are *properties* of the referent through which the subject grasps the referent.

These assumptions, together with the claim that *pain* is a referring expression, add up to the following argument:

1) A concept refers to an entity via a mode of presentation. The mode of presentation provides the route by which the entity is picked out by the concept.

2) Modes of presentation are properties of the referent.

3) If the same mode of presentation is associated with two (coreferring) concepts, it is knowable a priori that these concepts corefer.

4) The concept *pain* is a referring expression.

5) Physical or functional concepts have as their mode of presentation physical or functional properties.

6) There is no physical or functional concept such that it is knowable a priori that such a concept corefers with *pain*.

Lemma:

7) No physical or functional property of pain could provide the route by which pain is picked out by the concept *pain.*

8) Properties are either physical, or functional, or irreducibly mental.

---

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

Conclusion:
9) Pain has at least one property that is irreducibly mental.[74]

Notice that this argument does not beg the question against a posteriori identities. *Water* and H_2O can have their referent in common, since the same substance, water, has many different physical and chemical properties, and these properties can all serve as modes of presentation providing different routes to the referent. But in the phenomenal case there is a problem. If the concept *pain* referred via a physical or functional property, there would have to be physical or functional concept that would have the same mode of presentation as *pain* has. But that would have to be knowable a priori, according to premiss 3. This is not the case. In fact, it seems that any thought of the form

Pain is X,

[74]I present the argument in a somewhat simplified form which I hope preserves all the essential elements of the original. However, there is one difference. White uses the argument as a *reductio* for analytic functionalism. He attempts to show that the denial of analytic functionalism leads to Dualism, and he claims that that is unacceptable. (The same view is advocated in Smart (1959).) I treat this argument as an anti-physicalist argument because I think that analytic functionalism is extremely implausible. However, for a different argument for analytic functionalism, see Levin (1986).

where *X* stands for a physical or functional concept, is conceivably false.[75] Now Dualism follows since, if the mode of presentation associated with *pain* is irreducibly mental, then there are properties that are neither physical nor functional.

There are a number of problems with the argument. First of all, not everybody would share White's conception of modes of presentation. According to Fodor (1990), e.g., the mode of presentation of the concept *pain* is simply a Mentalese term; it is not a property of the referent, and it does not "provide a route" via which the referent is picked out by the concept. On Fodor's view, then premiss 1 and 2 are false. On this account it could be, e.g., that the Mentalese term 'pain' and the Mentalese term 'C-fibre firing', e.g., are both causally related in the appropriate way for reference to a physical property, without these terms being associated with different properties that provide the route to the referent.

The main problem with the argument is premiss 3. It is a relative of *CTT* in Descartes' argument, in that, while Descartes thesis asserted the transparency of clear and distinct perception of substances, it states that mode of presentation is transparent. I will call it *White's Transparency Thesis* (*WTT*).

---

[75]This claim is based on the same intuition as the claim that *zombies* are conceivable.

(*WTT*) If the same mode of presentation is associated with two (coreferring) concepts, it is knowable a priori that these concepts corefer.

This thesis provides the missing link for the BAT Argument and the Mary Argument. The question there was why the mode of presentation of *pain* and the mode of presentation of some physical or functional concept cannot pick out the same state (or, in Nagel's terms, why can we not form a physical conception of phenomenal states). *WTT* can be put to work to answer this question.

Notice that the mode of presentation of both *pain*, and of physical or functional concepts is an essential property of the referent. *Painfulness*, the mode of presentation of *pain*, is an essential property of pain, and the same is, presumably true of *Piramidal cell activity*, or any other physical or functional concept.[76] The mode of presentation of these concepts does not utilize contingent properties of the referent, like, e.g., the mode of presentation of *water* does. But then, if *pain* and some physical or functional concept referred to the same state, their mode of presentation would have to be the same. However, then, according to *WTT*, we would have to be able to tell a priori that *pain* refers to the same property as some physical or functional term does. But we cannot, so *pain* cannot refer to any physical or functional property.

---

[76]See Kripke 1972.

What is the justification for *WTT*? White offers a reduction in support of it. Suppose that two expressions are coreferential, and that this fact cannot be established a priori and has not been established a posteriori. Suppose further that there are *not* two different properties in virtue of which the two descriptions pick out the same referent. Then there is a possible world where speakers who are epistemically equivalent to us use these terms to refer to different objects. As used by these people, these terms must pick out their referents in virtue of distinct properties, because unlike our terms, theirs pick out different objects. But this contradicts our initial assumption that there are *not* two different properties that serve as the mode of presentation of these concepts.

The argument has a questionable step. Just because it is not known a priori that two terms pick out their referent via the same property, it does not follow that these two terms might pick out different things in some possible world. Even if we accept that modes of presentation involve properties of the referent, we might want to deny that these properties exhaust all there is to modes of presentations. Properties of the referent might be involved in the way concepts pick out their referents; but a certain *relation* between the concept and the referent might be also involved. Different concepts might latch on to the same property providing the route from concept to referent in different ways.

This means that concept individuation, and so a priori knowledge of coreference, might be more fine-grained than reference fixation. The Fregean assumption on which White's argument relies, that mode of presentation performs both functions at the same time, might be mistaken. Mode of presentation, then, in the sense in which it fixes reference, might not be transparent, in the sense that it might not be knowable a priori whether two concepts have the same mode of presentation. In that case it would not follow from the fact that it is not knowable a priori that two concepts corefer that there is a possible world in which they do *not* corefer. In fact, on the account of phenomenal concepts given in Chapter Two, we have a model of why *WTT* flounders on the phenomenal case. More on this in Chapter Four.

3.5 KRIPKE'S ARGUMENT FOR DUALISM

Kripke's (1972) argument for Dualism can be seen as another suggestion how to answer the question left open by the Mary Argument: it can be seen as an attempt to show when two modes of presentation cannot determine the same referent. Kripke develops Descartes' Conceivability Argument for Dualism, and like Descartes (and White), he is also building the argument around a perceived link between epistemology, metaphysics, and the conceptual realm.

Semantics is key to Kripke's argument, so let me turn to it briefly. Kripke's semantics grew out of an opposition to the then prevailing Fregean view. He suggested that no single entity can do all the work Frege assigned to sense. In particular, no single intension function can fix the reference of all concepts for all possible worlds. As Kripke (1972) and Putnam (1975) pointed out, the reference fixer associated with, say, 'water', picks out water in our world, but picks out XYZ on Twin Earth. The reference of many concepts is determined by different mechanisms in the actual world and in counterfactual possible worlds; in the case of rigid designators the reference fixing description picks out the referent in the actual world and the reference in counterfactual possible worlds is then fixed to be the same as it is in the actual world.

This picture presents a radical departure from Fregean semantics. In addition to Kripke's claims about two-tiered reference fixing for rigid designators, there are a number of further differences between his and Frege's account of meaning. First, Kripke does not think that reference fixers are part of the *meaning* of a concept. According to him, reference is all there is to meaning. He also does not think that reference is always fixed via description, and so, *pace* Frege, he does not think that every competent user of a concept has to know some reference fixing description associated with the concept. Someone, e.g., might acquire the concept via ostensive definition, or just forget the description that initially served to fix the concept's reference, or acquire the concept from a competent user without ever learning the reference fixing description. Moreover, Kripke seems to think that a description might serve as a reference fixer even if it is not even *true* of the referent.

But the most important point of difference is that according to Kripke, reference fixers do not determine the reference of a concept independently of the context in which it is used. *Water* means H_2O, but if our environment turned out to be Twin Earth, it would mean XYZ. However, for all these differences, it seems that Kripke's semantics retained just enough of the doctrine of the transparency of meaning to be able to formulate a

Conceivability Argument. We are going to see shortly in what sense Kripke is a heir of Frege's transparency thesis.[77]

[77]Kripke's reference fixers, in spite of the differences from Fregean sense, still play many of the roles associated with modes of presentation: they individuate concepts, and, at least in the case of those concepts that actually have them, they determine reference within a *context*.

There is a certain tradition in philosophy that, *pace* Kant, identified necessity, analyticity and a prioricity.[78] The tradition was swept aside by Kripke's arguments that the relations between these concepts are much more complex than this. Kripke's claims are based on his insights into rigid designation and the two-tiered nature of reference fixing. He pointed out that there is a large class of necessary statements whose truth is only knowable a posteriori. *Water is H_2O*, given that it is true, is necessary. This follows from the fact that both *water* and *H_2O* are rigid designators. However, it is not knowable a priori since, before we discovered that water is H_2O, our evidence did not rule out that water has a different chemical structure, for example, that it is XYZ.

Conceptual and metaphysical necessity also seem to be different since, e.g., *I am here now* is conceptually necessary but does not express a metaphysical necessity, and *There are infinitely many prime numbers* expresses a metaphysical necessity but does not seem to be true (or knowable) solely in virtue of the meanings of its constituent expressions.

For a priori truths, the justification often consists in conceptual matters. So, for example, I know a priori that all bachelors are unmarried since my mastery of the concepts *unmarried* and *bachelor* involve my

[78]For example, the positivists and especially Carnap (1955).

knowing this; similarly for my knowing that I am here now. But some would argue that not every a priori state of knowledge can be justified solely in virtue of conceptual matters; e.g. I know a priori that there are infinitely many primes. And I know a priori that I am now thinking that my dream of last night could have been dreamt by someone else.

The argument

Kripke, after prying apart metaphysical necessity, conceptual necessity, and a prioricity, presented an argument that actually depends on a link between them, and especially between conceivability and metaphysical necessity. He develops and modifies the Cartesian claim that conceivability is our guide to possibility (and so to necessity).[79] It will turn out on Kripke's argument that *certain propositions could only be true if they are conceptually true.* More precisely, he will make the assumption that two terms, both of which have essential reference fixers that we actually know, can only pick out the same things if the synonymy is transparent. The argument takes the form of a challenge for the physicalist.

---

[79]The classic exposition of his views on the mind-body problem is in the last part of *Naming and Necessity* (Kripke 1972). My reconstruction makes the argument a little more explicit than what the text warrants; obviously, the following is my, hopefully not too uncharitable, *interpretation* of Kripke.

Kripke's idea is the following. He wants to argue from the *prima facie* possibility of *pain is not C-fibre firing* to its real possibility. Now, there is an obvious problem here. *Water is not H_2O* appears just as possible, even though we know it to be necessarily false. As Putnam has pointed out,

> we can perfectly well imagine having experiences that would convince us (and that would make it rational to believe) that water is not H_2O. In that sense, it is conceivable that water isn't H_2O. (Putnam 1975, p. 233)

However, as we will see, we can explain away the apparent possibility that it is not. Now take a thought like *Pain is brain state C-fibre firing*. It is, again, *prima facie* possible that pain is not C-fibre firing, but in this case we cannot explain away this prima facie possibility. And, if we cannot explain away the apparent, or prima facie possibility of the falsity of an identity statement, we have identified a real possibility. So, concludes Kripke, pain is not C-fibre firing. This argument has the obvious advantage that it takes into account the existence of a posteriori identities in a way that the Mary Argument did not.

Let’s state the argument more formally:

1) It is prima facie possible that C-fibre firing is not pain.[80]

---

[80]The argument is formulated to refute the proposition that pain is C-fibre firing; but we can treat 'C-fibre firing' simply as a place holder for any physical or functional term at all.

2) If a state of affairs is prima facie possible, and there is no way of explaining away this prima facie possibility, then this state of affairs is (metaphysically) possible.

3) There is no way of explaining away the prima facie possibility of C-fibre firing not being pain.

4) By 1, 2 and 3, it is (metaphysically) possible that C-fibre firing is not pain.

5) But if it is (metaphysically) possible that pain is not C-fibre firing, then, since both *pain* and *C-fibre firing* are rigid designators, pain is not C-fibre firing.

6) By 4 and 5, pain is not C-fibre firing.

Obviously, the most controversial premisses here are 2 and 3. It is not even clear what they are saying. What does the notion of '*prima facie* possibility' come to? And what is it to "explain away" prima facie possibility? Let's first deal with the notion of prima facie possibility. On first approximation we can interpret it as conceivability simpliciter, as defined in the introduction:

(Con) A statement S is conceivable, if it is logically consistent with the totality of conceptual truths, i.e., if -S is not a conceptual truth.

In this sense it is also clearly conceivable that water is not H_2O.[81] The truth of *Water is H_2O*, just like the putative truth of *Pain is C-fibre firing* is not analytic. Just by grasping the meaning of the statement, you cannot come to know its truth.

---

[81] I.e., under the mode of presentation *Water is not H_2O*, and not under the mode of presentation *Water is not water*.

Yet it is not metaphysically possible that water is not H_2O. As Kripke famously pointed it out, since both terms are rigid designators, the identity statement is necessary, even though it is a posteriori. Prima facie possibility in itself then does not imply *real* (metaphysical) possibility. So Kripke has to come up with a way to distinguish between those cases of prima facie possibility that do not give us a reason to infer to real possibility (as in the water-H_2O case), and those that do (as Kripke claims is the case in the pain-C-fibre firing case).

Here comes in the idea of "explaining away" prima facie possibility. Kripke finds the difference between the two cases in the applicability of a certain strategy to explain away the apparent possibility of water not being H_2O, that is, to explain why certain statements appear to be contingent even if they are not. The explanation goes like this. When you entertain a proposition, e.g., that water is not H_2O, you entertain the proposition via a mode of presentation.[82] There may be many different modes of presentation

[82]There seems to be a tension in Kripke's (1972) views on meaning. As we have seen, one of the main aims of the book is to debunk the Fregean theory of meaning according to which there is a (presumably descriptive) sense or mode of presentation associated with each term that determines the reference of the term. Kripke proposed instead the theory of direct reference advocating a return to the earlier Millian idea that the only semantic contribution of expressions is their reference (cf. Mill 1843). At the same time, his argument for Dualism seems to rely on the notion of a reference fixer (a description sometimes associated with terms that determines their reference) which seems to be closely related to the notion of

expressing the same proposition in a given context, e.g., *water is not H_2O*, *H_2O is not H_2O*, *the substance that is actually grandma's favorite drink is not water*, etc. These modes of presentation correspond to the that-clauses used to express belief-attributions.[83]

According to Kripke, the concept *water* is the concept of a substance that is *actually* watery (i.e., the concept *water*, or rather, the mode of presentation associated with *water*, is a rigidification of the description 'clear liquid that fills the oceans, is tasteless, etc.'). This means that the mode of presentation *water is not H_2O* can be spelled out as the mode *the substance that is actually watery is not H_2O*[84]. Now, the explanation of why *Water is not H_2O* seems prima facie possible, according to Kripke, is that it is possible for someone to have the thought with the mode *the substance that

---

a mode of presentation. Indeed, what could conceivability come to if all there was to meaning was reference? A two-dimensional framework for reference fixation of general terms (and even, perhaps, for proper names and indexicals), as developed later in this chapter, might deal with many of the problems for the description theory raised by Kripke, and so might help reconcile Fregean and Kripkean semantics to a large extent. In any case, I am assuming that something like a notion of mode of presentation is what Kripke has in mind by his term 'reference fixer'.

[83]At least there is a rough correspondence. See Loar (1988) for the view that thoughts (or, what he calls 'psychological content') only roughly correlate with that-clauses.

[84]For simplicity, I have not tried to spell out the mode of presentation of the concept *H_2O*. All the important work here is done by the mode of presentation of *water*.

is actually watery is not H_2O* and express a *true* proposition via that mode. It is possible because there is a possible world where the stuff that is actually watery is not H_2O, but, say, XYZ.[85] So conceivability simpliciter is an indicator of possibility, but only an indicator that it is possible for a thought with the same mode to express a truth is some possible world, *considered as actual*. As things are, the thought *Water is not H_2O* is false in every possible world *considered as counterfactual*,[86] i.e., it is not really metaphysically possible for water not to be H_2O.

---

[85]Sometimes Kripke talks as if mode of presentation or thought, in addition to being descriptive, were internal. He explains the apparent possibility of water not being H_2O thus:

> although the statement itself is necessary, someone could, *qualitatively* speaking, be in the same epistemic situation as the original, and in such a situation a *qualitatively* analogous statement could be false. (p. 150)

This, however, suggests that Kripke identified modes of presentations (or reference fixers, as he likes to put it) with what has become known as "narrow content" in the literature: his reference fixers determine reference, are descriptive and they supervene on the intrinsic states of the agent. I cannot take up the issue here whether the notion of narrow content is a workable one; I will have a little more to say about this in connection with Jackson's and Chalmers' argument. But let me point out that narrow content does not seem to be required by the notion of conceivability we will be working with here. If modes of presentations were wide content descriptions, transparency could still hold.

[86]The phrases 'worlds considered as actual', 'worlds considered as counterfactual' are due to Davies and Humberstone (1980). It is a familiar notion from two-dimensional modal logic, developed in part on the inspiration of Kripke's account of a posteriori necessity, that thoughts about possible

Kripke's idea is that the above considerations help explain away why it seems possible that water is not H_2O even though it is not really possible. I think it is safe to interpret him as concluding that conceivability simpliciter is not a sure guide to possibility, it only indicates that there is a possible world such that, *were it the actual world*, a thought which is conceivable would express a truth. We can learn, like in the case of the thought *Water is not H_2O*, that it is not really possible for it to be true by looking at some contingent background truths *that fix the reference* of the concept *water*. Once we have learned that H_2O is the actual clear liquid that quenches thirst, etc., we will no longer think it possible for water not to be H_2O. Looking at the totality of contingent truths about the world which fixes the reference of our terms will help explain away any mere "prima facie" possibility.

What about the thought *Pain is not C-fibre firing*? The concept *pain* has the mode of presentation *the sensation that feels painful*. However, this mode of presentation is not, so to speak, accidental, as in the case of *water* and *the substance that is *actually* watery*. In the case of *water* the reference is picked out through an accidental feature of the referent; in the

---

worlds are evaluated differently depending on whether the worlds are considered as actual or considered as counterfactual. More on this in the next sections.

case of *pain* it is picked out through an essential feature of it: whatever feels painful just *is* pain, irrespective of the context.

So let's now suppose that pain *is* C-fibre firing. Can one explain away the prima facie possibility that pain is not C-fibre firing the same way we did with the prima facie possibility that water is not H_2O? In other words, can one entertain the mode *the sensation that feels painful is not C-fibre firing*[87], and express a true proposition in some world considered as actual, even if the thought *The sensation that feels painful is not C-fibre firing* is necessarily false? If pain is C-fibre firing, and so *Pain is not C-fibre firing* is not possible, the answer is obviously no. *The sensation that feels painful is not C-fibre firing* could only express a true proposition if *Pain is C-fibre firing* was false.

The reason is that in any possible world in which we think *the sensation that feels painful is not C-fibre firing* we are thinking the very same proposition. The mode of presentation *the sensation that feels painful*, as well as the mode of presentation *C-fibre firing*, fixes the reference *essentially*. So we cannot explain away the illusion of possibility by

---

[87]Again, I am not going to spell out the mode of presentation of C-fibre firing; however, I assume all along that whatever it is, it is *essential* rather than accidental. As Kripke says:

> ...'pain' is a rigid designator of the ...phenomenon it designates: if something is a pain it is essentially so...The same holds for the term 'C-fiber stimulation', provided that 'C-fibers' is a rigid designator....

saying that, while the thought *Pain (the sensation that feels painful) is not C-fibre firing* is true in some possible world considered as actual, it is neither conceivable *together with all the relevant background truths that fix its reference*, nor is it true in any possible world considered as counterfactual. In the case of *Pain is C-fibre firing* it does not make a difference whether you think about possible worlds as actual or as counterfactual. Kripke's claim is that the prima facie possibility of pain not being C-fibre firing cannot be explained away, and so, by the necessity of identity, pain *is not* C-fibre firing.

We can see now how Kripke's fills in the gaps in the Knowledge Argument. Our problem there was that to conclude from not K(Mary, *what it's like to see red is R-fibre firing*, that what it's like to see red is R-fibre firing) to the truth of *What it's like to see red is not R-fibre firing*, we would have to have a way of seeing that *what it's like to see red*, and *R-fibre firing* could not pick out the same kind of state. Kripke supplies an argument that, if it succeeds, shows just that. He argues that these modes cannot pick out the same kind of state because they are both essential reference fixers, and so the *prima facie* possibility of what it's like to see red not being R-fibre firing cannot be explained away appealing to these modes of presentations in the same manner that the *prima facie* possibility of water not being H_2O could be explained away.

But why is Kripke so certain that if the prima facie possibility of what it's like to see red not being R-fibre firing cannot be "explained away" on the same model as we explained away the prima facie possibility of water not being H_2O, it cannot be explained away at all? Why think that, just because it is conceivable that what it's like to see red is not R-fibre firing, there is even a possible world *considered as actual* such that the thought *What it's like to see red is not R-fibre firing* is true there? Perhaps conceivability is not a guide even to possibility in this weak sense?

The reason that it is possible for *Water is not H_2O* to be true in some possible world considered as actual is that one of the concepts, the concept *water* has a reference fixer that utilizes a *contingent* property of that referent (clear liquid, etc.), and so there are worlds where the reference fixing property is instantiated without **H_2O** being instantiated. But if both concepts flanking the identity sign have essential reference fixers, all bets are off.

The only reason to think that, in the case where both concepts have essential reference fixers, the mere conceivability of the failure of identity shows that the reference fixers are different, is to hold the following transparency thesis. I will call it *Kripkean Transparency Thesis* (*KTT*):

> (*KTT*) it is impossible to refer to the same kind of state through rigid designators with essential reference fixers that have no conceptual connection at all.

This is a variant of the Transparency Theses we have seen proposed by Descartes and White in their respective Conceivability Arguments. Even though much weaker that the transparency thesis advocated by Frege, it is strong enough to support a Conceivability Argument. The problem is, it is unargued for, and, we will see, can be undone by the very account of phenomenal concepts proposed in Chapter Two. Again, I postpone detailed discussion till Chapter Four.

3.6 THE NEW CONCEIVABILITY ARGUMENTS

The Conceivability Arguments we have been looking at so far target the two main physicalist accounts of mind: identity theory and functionalism. However, even if both the identity theory and functionalism is false, Physicalism still could be true if there was an alternative way to account for the truth of P. Jackson (1993, 1995) and Chalmers (1996) go further than most: they do not just attack identity theory and functionalism, they try to give general reasons why P is false. They are trying to show that *any* physicalist account of the mind is doomed to failure.

Jackson and Chalmers develops the semantic framework proposed by Kripke to give the Conceivability Arguments their most sophisticated and

well-articulated form. While among the Conceivability Arguments these pose the greatest challenge to physicalism, I will argue in Chapter Four that they ultimately fail. To show this, I am going to construct a master argument called the ‘Zombie Refutation’ that will refute Jackson and Chalmers’ argument. I will be able to generalize this argument to refute all other extant versions of the Conceivability Arguments, since the argument attacks the link between conceivability and metaphysical possibility on which they all rely.

Jackson's and Chalmers' arguments are similar. Their definitions of Physicalism are almost identical, as is the semantical framework in which they formulate their arguments. The main difference between their arguments if that they employ different formulations of the crucial premiss linking conceivability and possibility. But these premisses are still closely related. At the end of this section, I will show that Chalmers' main premiss implies Jackson's main premiss. This will be important, since then by refuting Jackson’s premiss I can refute Chalmers’ premiss as well. I will be explaining Jackson's argument first, but I will point out along the way the similarities and differences with Chalmers' argument.

3.6.1 JACKSON'S ARGUMENT

Before presenting the argument that I call "Jackson's argument" I should admit that Jackson never explicitly endorses this argument and recently he seems to reject its conclusion. Jackson himself presents his argument as a challenge for the physicalist, rather than a straight refutation of Physicalism. But if, as Jackson's own Knowledge Argument (Jackson 1982) assumes, and as all other proponents of the Conceivability Argument claim, phenomenal concepts do not have the requisite conceptual connections with non-phenomenal concepts, his argument can easily be turned into a refutation of Physicalism. He explicitly argues for a premiss that links metaphysical and conceptual possibility in a way that, given the unanalyzability of phenomenal concepts, makes the anti-physicalist conclusion almost inevitable.

Jackson's own view now seems to be that, since anti-Physicalism is so implausible, there must be something wrong with the Conceivability Argument against Physicalism; but he professes to be uncommitted as to what exactly is wrong with it. He calls this position the 'there must be a reply' reply (Jackson 1996, pp. 134-5). In what follows, I treat his argument as an anti-physicalist argument; but I do not want to make much of attributing the conclusion to him.

In a nutshell, Jackson's argument is the following[88]. 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. He then argues 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 support such a prioricities. It follows that if there are true phenomenal statements then Physicalism is false.[89] Since we are assuming that there are phenomenal truths, Physicalism is refuted. Let us now look at the argument a little more closely.

Physicalism

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

---

[88]My presentation is based on Jackson (1993) and (1995).

[89]Jackson does not explicitly take this last step, although he does not quite say which premiss he thinks might be false. One might take the view that phenomenal concepts may have functional or other analyses; he alludes to this possibility (ibid, p. 142). However, I want to put this view aside for two reasons. First, I find this view very implausible; second, on this view the Conceivability Arguments against Physicalism cannot be formulated. I want to grant to the proponent of the Conceivability Arguments as much as possible before I present my refutation; I do not want our defense of Physicalism to depend on such a contentious semantic doctrine.

P Among worlds where no property alien to the actual world is instantiated, any two that are exactly alike with respect to their complete descriptions (including specification of the fundamental laws) in the language of the ideal fundamental physical theory are duplicates simpliciter.[90]

In Chapter One we have seen how this definition makes more precise the intuitive idea underlying Physicalism: that there is nothing over and above the physical stuff in our world. Jackson also suggests that P is equivalent to the claim that every truth T about our world, be it physical, chemical, biological, psychological, etc., is necessitated by a statement of physics K that gives the full physical description of the world, together with the statement that K *is* the full, fundamental description of our world. This is the Entailment Thesis:

(E) For any true statement T,
$\Box(K\&C\&F\supset T)$,

which is equivalent to P.

If P is true, for most true statements T, $\Box(K\supset T)$ will be true. Those truths for which this does not hold are "global" truths, in that their truth depends on the global distribution of fundamental properties. It is clear that positive phenomenal properties, e.g., being in pain, are not global. So, for the purposes of the arguments to follow, we can use the simplified formulation

[90]Jackson actually gives a slightly different version of P; the differences, however, will not matter in what follows. Chalmers (1996), pp. 41-42, formulates physicalism in a very similar vein.

(E) For any true statement T,
$\Box(K \supset T)$,

keeping in mind that this is not strictly equivalent to P.

Conceptual explanation

According to Jackson, the necessities '$K \supset T$' cannot be brute facts; they need explaining.[91] Jackson observes that if T is, e.g., a psychological statement then analytical functionalism has a story to tell about why the statement is necessary. As he puts it:

> ....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. (p. 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". He thinks that the explanation has to be, in an appropriate sense, *conceptual.*

Jackson argues that if Physicalism is true then '$K \supset T$' is not only metaphysically necessary, but is also an a priori conceptual truth; i.e., he argues 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

[91]See Section 1.4 for more discussion on this.

A Priori Entailment Thesis:

(APET) if (E) is true, for any true T, statements of the form

$K \supset T$

are conceptual truths.[92]

The *APET* plays the same role in the argument as *CTT*, *WTT*, and *KTT*, i.e., the respective Transparency Theses played in Descartes', White's, and Kripke's argument: it links conceivability and possibility. It is a generalization of Kripke's thesis; it holds not only of identity statements, but covers the very supervenience claims that must be true if physicalism is true.

We have observed earlier that conceivability does not always imply possibility. As Kripke (1972) has pointed it out, identity statements containing rigid designators, if true, are necessarily true; e.g., water is necessarily H_2O. It is conceivable, however, that water is not H_2O. This shows that mere conceivability is not a reliable guide to possibility. Proponents of the new

---

[92]This, of course, is not an arbitrary requirement for Physicalism alone. Jackson claims that any metaphysical theory that makes a distinction between fundamental and non-fundamental properties, e.g., Berkelean idealism, or Cartesian Dualism, has to be able to produce, for any true T, appropriate derivations of the respective entailment claims

$K^* \supset T$,

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, e.g., the fundamental discourse is mentalistic, and all the physical truths have to be a priori entailed by a complete mentalistic description of the world.)

Conceivability Arguments claim to have identified a special class of statements for which conceivability *does* imply possibility. Jackson proposes that it is the class of statements which, *conjoined with the full truth about the world in the language of fundamental discourse,* are conceivable, so, if Physicalism is true, then it is the class of statements which, conjoined with the full physical truth about the world, are conceivable.

The *A Priori Entailment Thesis* is really a claim about the link between conceivability and possibility. We can see this by showing that the *APET* is equivalent to the

Conceivability-Possibility Thesis

> (C-P) if Physicalism is true then, for any thought *S*, if *K&S* is conceivable then it is possible.

Notice that *K&S* is conceivable iff *-(K&S)*, that is, *K⊃-S* is not a conceptual truth. Notice also that *K&S* is possible iff *□(K⊃-S)* is false. So substituting ‘it is not the case that *K⊃-S* is a conceptual truth' for ‘*K&S* is conceivable', and 'it is not the case that *□(K⊃-S)* for '*K&S* is possible', we get

> if Physicalism is true, then, for any thought *S*, if *□(K⊃-S)* is true, then *K⊃-S* is a conceptual truth,

which is just the *APET*.

Why think that the *APET* is true? Jackson provides the following

considerations. [93] 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 (Kripke 1972, pp. 140-162), Jackson observes that, arguably, in all *bona fide* cases of identity statements where the denial of the identity claim is conceivable (e.g., *Water is not H_2O*), there are contingent truths such that the denial of the identity statement, *in conjunction with them,* is not conceivable.

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.[94] 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. The full fundamental description of the universe always provides enough background information

---

[93]I am going to discuss a more technical, elaborate explanation in the next section.

[94]That it is a conceptual truth follows from Jackson's semantics. Jackson's semantics will be discussed in detail in the next section.

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 the 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

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, is a priori.[95] Since, according to Jackson, these contingent statements about H_2O are similarly a priori derivable, perhaps through some intermediary steps, from contingent truths of micro-physics, 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[96] can be similarly shown to be a priori entailed by the full physical description of the world.[97]

---

[95]Of course, the derivation, as it stands, is incomplete. To complete it, we would have to have the requisite conceptual truths that link the concept 'Earth', 'surface', 'clear', 'odorless', etc. to terms of lower level discourse, and, ultimately, micro-physics. But, according to Jackson, it is rather clear that such conceptual truths exist.

[96]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.

[97]One might think that, on the model of our derivation of 'Water covers 60% or the surface of the Earth', we can derive a priori, e.g., '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 a) and b) we get

d) Pain is C-fibre firing.

from a) and d) we get

e) (x has C-fibre firing & pain is C-fibre firing)⊃x has pain.

This derivation uses only contingent empirical truths and conceptual truths. It

Second, the *APET* is a very powerful explanatory claim. Modal claims of the form

$\Box(K \supset T)$

might seem metaphysically and epistemically mysterious. If correct, the *APET* 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

$K \supset M$

where K is the full fundamental description of the world, and M is any metaphysical truth, is a conceptual truth. 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 argument Jackson (and Chalmers) offer for the *APET* is: many putative necessities of the form

$\Box(K \supset T)$

---

has the form

$P \rightarrow x$ has pain,

where P is a conjunction of contingent facts of neurophysiology and psychology, and is knowable a priori. The problem with this derivation, however, is that one of the conjuncts in P, premiss 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.

do conform to the *APET*; there is no reason to suppose that there are exceptions to it[98]; and there are good explanatory motivations for it.

Jackson and Chalmers also supply a much more sophisticated and elaborate argument for the *APET*, based on the so-called two-dimensional semantic framework. They seem to suggest that the two-dimensional semantics, together with uncontroversial claims, entails the *APET*. However, I will show, they do not ultimately succeed in providing any additional support to the *APET*. Moreover, the defense for the *APET* in terms of the two-dimensional framework is not strictly necessary for an understanding of Jackson and Chalmers's statement of the Conceivability Argument, or my refutation of it. Hence, the reader can skip right through to the next subsection, the exposition of Jackson's anti-physicalist argument.

The two-dimensional account

[98]The main goal of this dissertation is to give such reasons. I will show that, contrary to Jackson, there *are* exceptions to the *APET*.

Jackson, as well as Chalmers, support their claim that conceptual truths play a key role in explaining *a posteriori* necessary truths[99] by applying two-dimensional semantics to mental representations; i.e., to concepts and thoughts.[100] On this account, concepts are *internally* individuated types of mental items; concepts are individuated independently of their referents so that we can meaningfully ask, for any world *w*, what the content of concept C would be, *were w the actual world.* According to the two-dimensional account, every thought or concept (or their linguistic counterpart) possesses *two* intensions - a primary intension and a secondary intension. The notion of primary intension is supposed to capture the internal aspect of content; i.e., it aims to capture, from an internal point of view, how the thinker *conceives* of the world. This notion presumably plays a key role in psychological explanation. The notion of secondary intension, on the other hand, is just another term for what we ordinarily, post-Putnam, mean by a concept's 'content'. For example, according to Jackson, the primary intension of *water* is, very roughly, *watery stuff*, i.e., *clear, odorless, etc...liquid around here that fills the oceans and lakes, etc.* The secondary intension of *water* is H_2O.

---

[99]Similar ideas have been formulated by Putnam (1975).

[100]This approach was first formulated by Kaplan (1978), (1979) and (1989), although he restricted the framework to demonstratives. It has been developed by Stalnaker (1978), Lewis (1979), Evans (1979), Davies and Humberstone (1980), and others.

To put it more formally, both the primary and the secondary intensions are functions from possible worlds to referents of the appropriate type; i.e., the primary and secondary intensions of thoughts are functions from possible worlds to {T,F}.[101] A concept's primary intension at a world **w** is evaluated without any reference to the actual world; the value of the primary intension function is determined by considering what the concept would apply to in **w,** *were* ***w*** *the actual world*. In this way, primary intension captures the internal determinants of content, since, in the absence of determination by the actual world "outside of the head", primary intension is supposed to be determined entirely by matters "inside the head". So, for example, the value of the primary intension of *water* in **w** is whatever stuff the concept *water* would apply to at **w** *were* ***w*** *the world that fixes the referent of the concept.* The primary intension of our concept *water* can be captured, according to Jackson, as *watery stuff,* i.e., *clear, odorless, etc....liquid around here that fills the oceans and lakes, etc.* If the primary intension of *water* is determined by something like the description *the clear, odorless, etc... liquid around,* then the value of the primary intension function at world **w** is whatever is the

[101]Strictly speaking, there is a difference in the arguments of these functions; *primary intension* is a function from *centered possible worlds* to referents, *secondary intension* is a function simply from possible worlds to referents. But for the purposes of the dissertation I will ignore this complication.

clear, odorless, etc... liquid at **w**. In the actual world that liquid is H_2O, but in some other world **w**, it - if it exists at all - may be different from H_2O, e.g., XYZ.

On the other hand, when evaluating a concept's secondary intension in a world **w**, one takes first the actual world and sees how it fixes the concept's referent, i.e., one first determines what the primary intension picks out as referent in the actual world, and then considers, in the light of that, what the concept would apply to in **w**. Secondary intension then coincides with the ordinary notion of content. It has an external element since what the referent is in counterfactual worlds generally depends on what the actual world is like. If Jackson is right that the primary intension of *water* is determined by the description *watery stuff*, then the secondary intension of our concept *water* is determined by a rigidified description, *dthat (watery stuff),* where *watery stuff* is the primary intension of *water*, and *dthat* is Kaplan's (1979) rigidifying operator, converting the primary intension into a rigid designator that picks out in every possible world whatever the value of the primary intension was in the actual world. The value of the secondary intension function of *water* in the actual world is the actual watery stuff, i.e., H_2O; and it is also H_2O in all other possible worlds, irrespective of what the local watery stuff is there. Notice however, that in the actual world the value of the primary and secondary intensions of a concept always coincide.

The two-dimensional apparatus outlined here is committed to two basic ideas as far as primary intension is concerned. One is the idea that there is a well-defined function from possible worlds to referents that is determined by what a concept or thought[102] *would* pick out as referent under different circumstances, taken as if they were actual. Let's call it the *Function Thesis*. The other is the idea that this function represents an aspect of content, usually distinct from ordinary content, which is also a function from possible worlds to referents. Let's call it the *Content Thesis*.[103]

There is a *prima facie* difficulty with the *Function Thesis*; it is not clear that the primary intension function has been well-specified. The problem is that it is hard to see how to evaluate the primary intension function in possible worlds where the concept or thought is not present. After all, the two-dimensional account defines primary intension as a function that is determined by what an (internally individuated) concept or thought *would* pick out as referent in different possible worlds, *were* these worlds the actual world. There might be no answer to this question in worlds where the

---

[102]Or their linguistic counterparts.

[103]There is an obvious sense in which the primary intension of a concept is determined just by what is in the head. The actual external features of the thinker's environment are irrelevant. For this reason it has seemed to some philosophers that the primary intension of a concept is its *narrow content*.

concept is not present. One might reply that we take the concept or thought from the actual world, and evaluate it in the possible worlds in question.[104] But how do we do this? After all, two-dimensional semantics is a theory that strives to give an account of what concepts *are*; it cannot take the notion of a concept for granted.

Jackson and Chalmers tries to solve this problem, as well as support their respective version of the *APET*, by introducing further, substantial commitments into their version of the two-dimensional account. One way to interpret the primary intension function in worlds where the concept does not exists, as suggested by both Jackson and Chalmers, would be to characterize primary intension as determined or constituted by a description associated with each concept (e.g., '*clear, odorless...etc. liquid...*'). These descriptions then pick out the appropriate referent of the concept in each possible world *considered as actual* (e.g., H_2O in the actual world, XYZ in Twin-universe, etc.); thereby constituting a mapping between worlds and referents. Let's call it the *Description Thesis.*

[104]Chalmers takes this view on p. 60, fn 26. He says "I think the primary intension is naturally extendible to a wider class of worlds: we can retain the concept from our own world, and consider how it applies to other worlds considered as actual..."

But this is a problematic assumption. Since, as we have seen, the actual world[105] plays no role in the determination of primary intension, primary intension must supervene on what is "in the head"; that is, it must be narrowly constituted. But it is quite controversial whether the content of the descriptions claimed to account for primary intension is narrowly constituted, or can be constructed out of narrowly constituted descriptions.[106] Even if some descriptions, like, e.g., phenomenal descriptions, are narrowly constituted, it is not clear that all of the requisite descriptions can be ultimately reduced to such narrow descriptions. One way to circumvent this problem is, instead of looking at the alleged conceptual truths one by one, to rather look at them as a network. Ramseyfication of our complete conceptual net might provide us with conceptual analyses that are both narrow and capable of determining primary intension. However, it is far from clear whether this suggestion can be carried out, and many philosophers object to the idea that there are enough non-trivial conceptual truths for the analyses to succeed (cf. Block and Stalnaker 1997). But I want to put this problem aside for now.

---

[105]Perhaps more accurately, the actual world *minus* what is "in the head" does not play any role in that.

[106]For an account of this strategy, and an indication of why it might not be viable, see Loewer (1985).

Another way to approach the problem, and this is the way that both Jackson and Chalmers seems ultimately to prefer, is to say that possessing a concept means that one has an a priori ability to tell, given some (presumably fundamental) description of a world, what (if anything) the concept would apply to, *were that world the actual world.* According to this proposal, given any fundamental description of any possible world, one will be able to figure out whether, were that the actual world, there would be any water, trees, spiders, consciousness, etc. there.[107] Let's call this the *A Priori Availability Thesis (APAT)*.

Both the *Description Thesis*, and the *APAT* are quite substantial theses, and they go beyond the basic two-dimensional framework, that is, the *Function Thesis*, and the *Content Thesis*. They try to capture the idea that primary intension is not only *specifiable for the subject* from a third person point of view; it is also *accessible to the subject* from the first person point of view.

---

[107]This, of course, is an idealization; what Chalmers and Jackson have in mind is what an ideal logician, not being bound by time constraints, powers of physical endurance, etc. could figure out under optimal circumstances.

It is also arguable that this thesis presupposes the *Description Thesis*. How else would one be able to compute the references of concepts in different possible worlds *considered as actual*? But I leave this question for now.

We have seen how the *Description Thesis* and the *APAT* provides support to the *Function Thesis*. Without accepting either of them, there is little reason to subscribe to the *Content Thesis* either, even supposing that the primary intension function exists. It is not clear that a theory which does not subscribe to the a priori availability of primary intension, and this includes most plausible accounts of content, could count primary intension as an aspect of *content*, as opposed to just a mathematical construction that has no particular relevance for semantics or psychology. On some theories of content (e.g., Fodor 1990, Chapter 4), the concept *water* might come to refer to just about anything in another possible world, given that things are set up right in that world. It is hard to see how, on this view, primary intension can be a variety of *content*.

However, there are many problems with both the *Description Thesis*, and the *APAT*. I already mentioned the difficulties of working out the *Description Thesis*. As for the *APAT*, I only want to point out here that the key role that Jackson and Chalmers assign to the two-dimensional semantics, and more specifically, to the *APAT,* in justifying the *APET*, is dubious. The *APAT* is supposed to provide support to the *APET*. However, while it is both true that the *APAT* is sufficient, in itself, to establish an anti-physicalist conclusion, and that the *APET* does follow from the *APAT*, it also

seems clear that the *APAT* stands just as much in need of a justification as the *APET* does.

Let us see first how Jackson uses the *APAT* to argue for the *APET.* The *APAT* says that there is something about our concepts that enables us to know a priori how the way the actual world turns out determines their referents. This does not just mean that we know a priori that the concept *water* refers to water. This knowledge is quite trivial. Rather, in the case of *water* we seem to know a priori roughly that water is whatever is the local clear, odorless, etc.... liquid that fills the rivers and oceans, etc. That is, possessing the concept *water* enables one to know a priori how contingent facts about the actual world figure in determining what *water* refers to. It follows from this that, once we are given the full fundamental description of the actual world, we will know what *water* refers to, without any further enquiry.

It also follows, unsurprisingly, that *before* we are given the information about the actual world, in some sense we do not really know what our concept refers to, even though we understand the concept. Jackson says that the sense in which we do not know what our concept refers to is that, given a full description of a possible world in the language of fundamental discourse, we still do not know how to locate the reference of our concept there. Even if we have a full fundamental description of Twin Earth

universe,[108] we will not know, without sufficient information about the actual world, whether there is water there, since we do not know which substance is water in the actual world. What we do know a priori, by virtue of possessing the concept, is that there is something there (namely, XYZ) which, *were Twin Earth world the actual world*, would be the referent of our concept *water*.

This in turn, according to Jackson, helps explain how we can understand necessary statements without knowing them a priori. Understanding certain statements - on Jackson's view, most statements - does not require that we know their ordinary truth-conditions (i.e., their secondary intension), in the demanding sense explicated above. In other words, understanding a statement does not require that, given a full description of a possible world in the language of fundamental physics (or, more generally, in the language of the fundamental discourse), we should be able to tell whether it is true in that world. The reason, of course, is that, to evaluate the truth value of a statement at a world described to us, we need to know the references of its constituent terms *at the actual world*, but to *understand* the statement all we have to know is how the actual world determines its reference. This explains why we might not know a priori that water is H_2O, though we understand the claim that water is H_2O.

---

[108]The world I have in mind is Twin Earth universe, taken as distinct from the actual world.

What this means, according to Jackson, is that the only reason why we sometimes do not know certain necessary truths a priori, even when we understand them, is that we lack the contextual information that determines their truth-conditions.[109] If *Physicalism* is true then, the *APAT* tells us, this cannot be the case for statements of the form

$K \supset T$,

where K, as we have seen, includes the full physical description of the world, and T is any (contingent or necessary) truth. If we understand the statement, we have to know a priori that it is true[110], since, given Physicalism, we have all the contextual information we will ever need.

This is Jackson's justification for the *A Priori Entailment Thesis*, i.e., the claim that, given the truth of Physicalism, truths of the form

$K \supset T$

will always be knowable a priori. As we will see shortly, the *APET* then can be used to refute Physicalism.

But there is a shorter route to Dualism *via* the *APAT*. The *APAT* is sufficient in itself to establish an anti-physicalist conclusion with respect to

---

[109]This argument, incidentally, is a version the argument Kripke gives in (1972), pp. 142-43. Kripke, like Jackson, meant to use the argument to pose a challenge to Physicalism.

[110]Indeed, that it is *necessarily* true.

qualia. For, if we accept the *APET*, we would have to conclude that in a physicalistic world no qualia properties are instantiated. That means, since qualia properties *are* instantiated in our world, that the full physical description of our world could not be the full *fundamental* description of it. That, however, begs the question against Physicalism: it is quite dubious that in the case of *qualia* properties, we should be able to tell a priori, whether in a purely physical universe they are instantiated or not.

The important point to note is this. The basic two-dimensional framework, defined only by the *Function Thesis*, and the *Content Thesis*, can be held independently of the *Description Thesis*, and *APAT*. For example, Fodor (1987, Ch. 2), and Stalnaker (1978), (1990) proposes something like this.[111] But on this, non-tendentious interpretation, the two-dimensional semantics does not provide support for the *APET*. With the addition of the *Description Thesis*, and *APAT*, the *APET* does follow, but the *APAT* itself is just as contentious as the *APET* itself.

One last point. If *APET* is true, it has to be necessarily true, and so, since it is hard to see what the full physical description of the world could have to do with it, presumably a priori true. However, it does not seem to be

[111]Of course this means that another solution for the problem with the *Function Thesis* mentioned above has to be found.

a priori knowable. There seems to be nothing incoherent about the possibility of a world *APET* fails, yet the entailment is simply underwritten by metaphysical reduction (via identity or realization) between higher level and basic physical concepts. again counts against the derivability.

However, I would like to bracket these problems for the moment. For all these problems, *APET* could be true. The refutation of the Conceivability Arguments, and, ultimately, of all the Transparency Theses, including *APET*, will be the subject of the last chapter.

The Argument

I now want to show how the *APET* can be used to argue that Physicalism is false. If the *APET* is true, the physicalist faces trouble *vis a 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.

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 thought linking the description to a concept of a lower level theory, and ultimately to a concept of micro-physics. Now consider the claim

K⊃x feels pain.

To derive *x feels pain* a priori from K, there must be some conceptual truth connecting *pain* with a *non-phenomenal* description such that satisfaction of the description is *a priori* sufficient for *feels pain*. But, arguably, there are *no* such conceptual truths.[112] For any such non-phenomenal description we can *conceive* of its being satisfied without anyone feeling pain. *Pain* is, as we

---

[112]One might think that, on the model of our derivation of *Water covers 60% or the surface of the Earth*, we can derive a priori, e.g., *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 truths of neurophysiology and psychology, is knowable a priori. The problem with this derivation, however, is that one of the conjuncts in P, premiss 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.

have discussed in detail in Chapter Two, a direct recognitional concept; we do not apply the term, at least in our own case, on the basis of any evidence, sensory, behavioral, or physical, distinct from what the term picks out, i.e., distinct from the experience itself. *Pain* refers to **pain** directly, or rather, via an essential feature of it, say, painfulness.[113] But it follows from the *APET* that if *x feels pain* cannot be derived *a priori* from K, then

$$\Box(K \supset x \text{ feels pain})$$

is false, and so if *x feels pain* is true[114], then Physicalism is false.[115] To put it more formally:

1) If Physicalism is true, then for any true T, statements of the form
   $K \supset T$
   are conceptual truths.
2) There are some true statements Q to the effect that phenomenal conscious experience occurs (eliminativism about phenomenal experience is false).

---

[113]In fact, on a Kripkean direct reference theory, this applies to proper names, demonstratives, natural kind terms, etc. The point is that on the *Jackson-Chalmers* view, this feature is unique to phenomenal concepts.

[114]Another way to block his argument is to deny that there are phenomenal states; see, e.g., 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.

[115]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 *APET*. 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, e.g., red phenomenal experience occurs.

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

So

4) Physicalism is false.

Another, perhaps more intuitive way to formulate the same argument is in terms of conceivability. We have shown earlier that

(C-P) if Physicalism is true then, for any statement S, if 'K&S' is conceivable then it is possible

is equivalent to the *APET.*

Now the argument can be run like this. Add to (C-P) the claim that

(Z) 'K&Z' is conceivable,

where Z is the claim that there are zombies; i.e., that it is conceivable that there is a world physically exactly like ours, but where all creatures are zombies. It follows from (C-P) and (Z) that the zombie-world is possible; which means, on the assumption that *we* are not zombies, that Physicalism is false. For the rest of the dissertation , for expository reasons, I will stay with the earlier formulation.

3.6.2 CHALMERS' ARGUMENT[116]

Chalmers (1996, pp. 65-123) endorses Jackson's conclusion. He has made various claims in its support, and believes he has a refutation of Physicalism, although he never explicitly formulated the argument for it in his book. His various claims and assumptions can be put together to form an argument; this is what I am trying to do here. As we have noted, Chalmers' formulation of Physicalism, and his semantics are essentially the same as Jackson's. However, his crucial premiss is a bit different. In this section I am going to give a sketch of Chalmers' argument, and then show that its main premiss entails the main premiss of Jackson's argument. This will enable me to refute both arguments by just refuting Jackson's main premiss.

Chalmers builds his argument around the following claim:

Necessity-Contingency Thesis

(NCT) the primary intension of a necessary a posteriori thought must be contingent.

[116]The contents of this section are, again, fairly technical and not strictly necessary for the understanding of the main argument that follows. Readers can skip strait Chapter Four, the refutation of the Conceivability Arguments.

The primary and secondary intension of a thought both express a proposition, and these sometimes differ. This is an extension of the notion of primary and secondary intension for concepts. E.g., the primary proposition of *Water is wet*, is, roughly, the proposition that the clear, odorless,....etc. liquid that fills the oceans around here is wet,[117] whereas its secondary proposition is the proposition that H_2O is wet.

The *NCT* follows from the claim that if the primary proposition of a thought is necessary, then the thought is knowable a priori.[118] This, like the *A Priori Entailment Thesis*, is a consequence of the *A Priori Availability Thesis*. The *APAT*, as applied to thoughts, says that, given the full fundamental description of a possible world, one can know a priori, for any thought one understands, whether the thought were true in that world *were that world the actual world*. So, for example, if one were given a full fundamental description of Twin Earth universe,[119] one could know a priori that, were Twin Earth universe the actual world, *Water is wet* would be true. This

---

[117]This, of course, on the assumption that something like the *Description Thesis* is correct, which, as we have mentioned, is doubtful. But as a rough characterization, this will do.

[118]On p. 69 Chalmers writes: "The class of [....]truths [whose first intension is necessary] corresponds directly to the class of a priori truths."

[119]Twin Earth universe is a possible world where, instead of water, there is XYZ on Earth.

means that if the *APAT* is true, one is able to know a priori whether the primary proposition of a statement is necessary or contingent.[120] But then, if the primary proposition of a thought is necessary, since it is also knowable a priori that the primary and secondary propositions of any thought coincide in the actual world, the thought will be a priori. That is, if the primary proposition of a thought is necessary, the thought is knowable a priori; which is equivalent to the *NCT*.

Now we can construct an anti-physicalist argument from Chalmers's principle in the following way. Observe that if Physicalism is true then the thought $K \supset Q$, for any true phenomenal thought Q, is necessarily true. The thought $K \supset Q$ is quite clearly a posteriori. If follows from the *NCT* then, that its primary intension expresses a contingent proposition. But can the primary proposition of $K \supset Q$ express a contingent proposition, given our assumption that Physicalism is true?

[120]On the assumption, of course, that the range of possible worlds is knowable a priori; i.e., that the full fundamental description of any possible world is a priori available.

The primary and secondary proposition of K coincide. It is arguable that basic physical concepts are picked out by essential properties of the referent; especially on the assumption that basic physical concepts are functional. If, e.g., electrons are necessarily whatever plays the electron role, then *electron* will have the same primary and secondary intension.[121]

If the primary and secondary intension of K coincides, then the primary intension of $K \supset Q$ is $K \supset Q^{-}$.[122] But the primary intension of a true thought also expresses a true proposition, so if Q is true, Q^{-} must be true as well. From the definition of Physicalism it follows that if Q^{-} is true and Physicalism holds then $K \supset Q^{-}$ is necessary. So it is not contingent. Consequently, on the assumption that the *NCT* is true, and that there is a true phenomenal thought Q, Physicalism is false.[123]

---

[121]Chalmers considers the possibility that physical concepts work more like *water* in that they are rigidified descriptions. So, the concept *electron* would be roughly equivalent to *dthat(the entity that plays the electron role)* (p. 135). In this case the argument would not go through as we will see shortly. But I will ignore this possibility for now; what I ultimately want to show is that, even if the primary and secondary intensions of physical concepts coincide, the argument fails.

[122]For any thought S, S^{-} will stand for a thought expressing its primary proposition.

[123]See the appendix for a more precise, formal exposition of this argument.

Here is how Chalmers' *NCT* and Jackson's *A Priori Entailment Thesis* are related. The *APET* claims that

> if Physicalism is true then, for any true statement T, the statement
> K⊃T
> must be a priori.

The *NCT* says that

> the primary intension of a necessary a posteriori statement must be contingent.

I am going to show now, along the same lines that Chalmers' anti-physicalist argument proceeded, that the denial of the *A Priori Entailment Thesis* entails the denial of the *NCT*, which is just to say that the *NCT* entails the *APET.* Suppose that, contrary to the *APET*, Physicalism is true, but $K \supset T$, where *K* is the true, complete physical description of the world, and *T* is a true thought, is only knowable a posteriori. Can the primary intension associated with $K \supset T$ be contingent, as the *NCT* requires? The primary intension of $K \supset T$ will be some statement $K \supset T^{-}$, where T^{-} is the primary intension of *T* (the primary intension of *K*, again, is just *K*). But if *T* is true, T^{-} is true as well, so $K \supset T^{-}$ will have to be necessary, given our assumption that Physicalism is true. In other words, we have shown that the denial of the *APET* entails the denial of the *NCT*.

This result enables me to refute both arguments at the same time. If I can show, as I shortly will, that Jackson's principle is false, then, since it is

implied by Chalmers's principle, I have thereby shown that Chalmers's principle has to be false, too. But before I turn to the refutation of the Conceivability Arguments, I would like to discuss an argument which, though it does not rely on the same semantic considerations as the Conceivability Arguments do, is closely related to them: the Explanatory Gap Argument.

3.7 LEVINE'S GAP ARGUMENT

The Conceivability Arguments discussed in this chapter rely on a version of what I called the Transparency Thesis (except the BAT Argument and the Mary Argument, which, as we have seen, stand in need of one). The first three of them (Descartes', White's, and Kripke's) state, roughly, that whenever two concepts have the same reference fixer, it is knowable a priori that they do. The corresponding theses in the new Conceivability Arguments of Jackson and Chalmers generalize this idea from coreferring terms to metaphysical supervenience theses. The Transparency Theses are all rooted in a broadly Fregean semantics according to which mode of presentation both fixes the reference and individuates concepts at the same time. This presupposes that there are many conceptual truths. I do not see that the "old" Transparency Theses (*CTT, WTT,* and *KTT*), on any straightforward interpretation, *entail* the *APET*, so it is not clear exactly how many conceptual truths there have to exist for the "old" Transparency Theses to hold . But the *APET*, and, accordingly, the *NCT* clearly requires that there are very many of them, in fact sufficiently many so that they ground the *APET*. On the Jackson-Chalmers view, e.g., the truth of *Water is* H_2O is satisfactorily explained by conceptual truths, together with contingent basic physical truths.

We saw that exactly what conceptual truths are required and how they supposedly ground the *APET* is not completely clear.[124] Jackson seems to hold a kind of descriptivism on which most concepts have a descriptive sense that analytically ties them to other concepts. For example, *water, clear liquid, oceans* are connected by the belief that *Water is a clear liquid (at the usual temperatures) that fills the oceans etc.,* and *liquid*, *substance*, and *flow* are connected by the belief that *Liquids are substances that flow, ...*and so on. Jackson thinks that given the myriad of these analyticities we will be able to derive any non-fundamental truth from the full true fundamental description.. If the *APET* is correct, and if phenomenal statements cannot be derived in this way from the full physical description of our world, then it follows that the full physical description of our world is not the full fundamental description. There is more in heaven and earth (specifically in our minds) than physicalism dreams of!

We have seen that this account of the relationship between conceivability and metaphysical possibility is motivated by Kripkean examples, and provides an attractive if audacious picture of the metaphysics and epistemology of metaphysical modality. But we have also seen that the arguments for the *APET* are far from decisive. Given its reliance on the

[124]Recall that the primary intension of a concept is presumably determined by the analyticities involving that and related concepts.

analytic/synthetic distinction, the Jackson-Chalmers argument will doubtlessly be greeted in some quarters with an incredulous stare.

For philosophers who reject the analytic/synthetic distinction, the *a priori* entailment thesis will appear to be obviously mistaken.[125] And even some of those who think that the distinction is coherent will not think that there are nearly enough analyticities to sustain the *a priori* entailment thesis.[126] After all, it is not all that plausible that the alleged examples of analyticities Jackson and Chalmers appeal to, e.g., *Water is the clear liquid that fills the oceans...* are analytic. Couldn't it be discovered that in fact it is not water in the oceans and that water is not clear (we were tricked by evil demons into thinking so).

I do not mean that it is metaphysically possible that water is not a clear liquid. That, of course, is correct, but not at issue. What I mean is that someone could be competent with the concept *water*, i.e., use it to refer to water, and yet not believe or even disbelieve that *it is a clear liquid that fills*

---

[125]The *locus classicus* for arguments against the analytic/synthetic distinction is Quine (1951). While Quine's discussions are very famous and influential, it is a curious fact that there is little agreement concerning exactly what Quine's arguments are and, although many philosophers pay lip service to these arguments, there isn't much sign of their giving up the analytic/synthetic distinction. An exception is Fodor (1994, 1997) who rejects the distinction on the basis of an atomistic account of concept individuation.

[126]Block and Stalnaker (1997), e.g., attacks the conceivability arguments on this ground.

the oceans ..etc.[127] I will call the view of concepts on which there are no, or very few, analyticities the "Quine-Fodor account" (Q-F account).[128]

For philosophers who hold the Q-F account of concepts, the Jackson-Chalmers Conceivability Argument is a non-starter. But one has the feeling that the physicalist cannot escape the intuitions behind the conceivability arguments so easily. The question then arises of how physicalism and the relation of phenomenal facts to physical facts look if the analytic/synthetic distinction is rejected.

We saw that physicalism requires the existence of supervenience bridge principles of the form $\Box(K \supset T)$. On the Jackson-Chalmers view these

---

[127]This is Fodor's view. According to him it is metaphysically possible for someone to possess the concept *water* without having any of the beliefs that are supposed to be analytic involving the concept. There could even be, metaphysically speaking, a mind whose sole concept is *water* and has no beliefs. Of course, this does not mean that either of these (alleged) possibilities are nomologically possible. As far as anything Fodor says, it may be a matter of law that no one can have the concept *water* without believing that water is the clear liquid that fills the oceans....

[128]Quine goes much further than Fodor, since he not only rejects the analytic/synthetic distinction, but also apparently rejects the view that reference is a substantive language-world relation. He holds a deflationary account of reference and truth. Quine also rejects the coherence of the notion of metaphysical necessity that is central to the very formulation of physicalism. I do not see that this rejection follows from rejection of the analytic/synthetic distinction. In any case, it is an interesting question how one should formulate physicalism, a doctrine Quine avows, without reference to metaphysical necessity.

principles are accounted for by being conceptual truths (or rather being entailed by conceptual truths). Is it possible to make sense of the supervenience principles and physicalistic reduction on the Q-F account of concepts? The answer is surely affirmative.

Consider, for example, the reductive identity *Water is H2O.* The reason that we believe it is that it *explains* statements like *Water dissolves sugar*, *Water expands when frozen*, *Water is a clear liquid*, etc. These are not conceptual truths, but rather are central and well confirmed beliefs. On the Quine-Fodor view, in contrast to the Jackson-Chalmers view, the order of explanation is reversed. Where Jackson and Chalmers "explain" *Water is* H_2O by deriving it from fundamental physical truths and conceptual truths, on the Q-F view *Water is* H_2O, together with physical truths explains why water is a clear liquid, etc. So far as I can see there is no incompatibility between the Q-F account of concepts and physicalism.

But it is difficult to formulate a version of the conceivability argument on the the Q-F account of concepts. The trouble is that among the central and well confirmed beliefs we have involving phenomenal concepts are those that connect them causally to behavior, stimuli, other mental states. For example, *Headaches cause people to take aspirin*. But we have independent reasons to think that the causes of behavior are physical; in this case some neurophysiological state. The identification of headaches with

that state, or kind of state, will then explain why headaches cause people to take aspirin (or rather be part of the explanation). Thus it seems that *Headache is h-fibre firing* and *Water is H_2O* are on a par; at least so far.

Yet there seems as though there is an important difference between *Headache is h-fibre firing* and *Water isH_2O*. Joe Levine (1998) has tried to get at the difference via the idea of the *explanatory gap*. Levine observes that sometimes we seek an explanation for an identity claim. For example, consider the claim that a diamond is a highly compressed lump of charcoal. On first hearing this one might very well ask: “How can that be, since diamonds are hard and brilliant, while charcoal is soft and black?”. The question is not a request for a justification of the claim (though one might be requesting that as well) but rather a puzzlement that such different properties are co-instantiated. But once we learn certain relevant empirical information concerning the behavior of carbon atoms under extreme pressure, that they are hard and brilliant, we no longer are puzzled.

Levine thinks that the case is different for claims like *Headaches are h-fibre firing*. In this case we are puzzled, and no further information, at least none of the sort that anyone has ever proposed, dispels the puzzlement. Notice that learning that headaches are the cause of h-behavior and h-fibre firings are the cause of h-behavior does not dispel the puzzlement although it entails the identity *Headaches are h-fibre firing.*

Levine calls an identity claim that admits of an intelligible request for explanation a "gapy identity."[129] Although he is not clear on this point, I think he would count both *Diamonds are compressed charcoal* and *Headaches are h-fibre firing* (supposing it is true) to both be gapy identities though with the former the gap has been closed by further empirical information. On the other hand, *Diamonds are diamonds* is not a gappy identity. It is unintelligible to ask why it is so. But he nowhere offers a general account of "gapy identities."

Next Levine defines the notions of "thin conceivability" and "thick conceivability". He says that a situation S is "thinly conceivable" relative to a representation R of S, just in case S is conceptually possible relative to R. This means that R is consistent with conceptual truths. A situation S is "thickly conceivable relative to R" iff S is thinly conceivable relative to R, and any derivation we can construct from R to a formally inconsistent representation R' must include gappy identities.

Levine thinks that *Water is not H_2O* is thinly, but not thickly conceivable. On the other hand, he thinks that *Headaches are not h-fibre firing* (and any negation of an identity between a qualia concept and a physical or functional concept) is thickly conceivable. Levine does not

[129]Levine (1998), Chapter 5.

actually spell out the derivation of a conceptually inconsistent statement from *Water is not H_2O*, but presumably he has in mind employing premises like *Water is the clear liquid that fills the oceans and...* and *H_20 is the clear liquid that fills the oceans...* It is not completely obvious to me that a complete explanation "from top to bottom" can always be given along these lines, but I will let that pass for now. Let's suppose he is correct.

Why then is *Headache is not h-fibre firing* thickly conceivable? Levine's answer is that the best explanation of this is that it is metaphysically possible. In other words, *headache* and *h-fibre firing* refer to distinct properties. Let's call this *Levine's Principle (LP)*

(LP) Thickly conceivable situations are metaphysically possible.

This lead him to construct the following anti-physicalist argument:

1) If physicalism is true, then for any true statement T, $\Box(K \supset T)$.

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

3) It is thickly conceivably that $K \supset Q$ is false.

4) Thickly conceivably situations are metaphysically possible.

5) Physicalism is false.

Levine claims that this argument is plausibly sound (though he does not go so far as to endorse it) and that it is plausible even if one holds the Q-

F account of concepts that rejects analyticities. I am not so sure that this latter claim is correct. It depends whether he can spell out the distinction between non-gapy and gapy identities in a non question-begging way. But he never really explains exactly how to make this distinction, other than relying on our intuitions concerning when a request for an explanation is intelligible.

The other worry I have about Levine's argument is that it is not really clear to me that one can derive all the supervenience principles not involving phenomenal concepts required by physicalism from the full physical description and non-gapy identities. Levine has not shown how we can make the derivations. And if we cannot then his argument will prove too much. I will make it too easy to defeat physicalism. Going further with these objections requires clarifying the non-gapy/gapy distinction. I will leave that to Levine. But it really does not matter how he clarifies it for I will show in the next chapter that no matter how the distinction is drawn, the Gap Argument is not sound.

CHAPTER FOUR
THE ZOMBIE-REFUTATION

The dualist conclusion of the Conceivability Arguments is rather implausible on several counts. There are powerful reasons to believe that Physicalism is true. Moreover, the dualist position has some internal problems of its own.

First of all, it has to account for why psycho-physical 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. None of these options are very attractive. Chalmers seems to prefer epiphenomenalism but that would make it completely mysterious how we know about our own phenomenal states.[130]

Third, although the new Conceivability Arguments for Dualism rely solely on the conceivability of worlds exactly like ours physically, but lacking any phenomenal properties instantiated, and not on the converse, i.e., the conceivability of worlds exactly like ours phenomenally, but lacking in any physical properties instantiated, it appears that an advocate of the Conceivability Argument would have to condone the existence of purely

---

[130]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.

phenomenal worlds.[131] It is barely intelligible what a world like that would be like. Fortunately, it can actually be shown that the arguments for Dualism we have been considering are unsound.

In this chapter I will to show that the Conceivability Arguments, as well as the Gap Argument, fail. The reason they fail has to do with the very nature of phenomenal concepts that give rise to the conceivability of zombies. First I am going to present a rather powerful reason for doubting the soundness of these arguments. But we can do better than that. In the second step, I will formulate a definitive refutation of the Conceivability Arguments.

4.1 CONTEMPLATING THE TRANSPARENCY THESES

As we have seen, each of the Conceivability Arguments rely on their particular version of the Transparency Thesis. However, on the account of phenomenal concepts we have given in Chapter Two, it can be seen that there is no a priori reason to hold the Transparency Theses (as well as *Levine's Principle*, which is not, strictly speaking, a transparency thesis). The theses in question are the following:

---

[131]Descartes actually did in the *Meditations* (John Cottingham; Robert Stoothoff; Dugald Murdoch 1984).

Cartesian Transparency Thesis

(CTT) When you clearly and distinctly perceive of substance A and substance B, that is, you conceive of them *through their whole essence*, and still do not see that they are *the same substance*, then they must be different.

White's Transparency Thesis

(WTT) If the same mode of presentation is associated with two (coreferring) concepts, it is knowable a priori that these concepts corefer.

Kripkean Transparency Thesis

(KTT) it is impossible to refer to the same kind of state through rigid designators with essential reference fixers that have no conceptual connection at all.

A Priori Entailment Thesis:

(APET) if (E) is true, for any true T, statements of the form
$K \supset T$
are conceptual truths.

Necessity-Contingency Thesis

(NCT) the primary intension of a necessary a posteriori thought must be contingent.

Levine's Principle

(LP) Thickly conceivable situations are metaphysically possible.

However, there is no a priori reason why the following scenario could not happen. We could have[132] two psychologically different conceptions (say, a basic recognitional concept, and a neurophysiological concept) of the same referent, that are connected to their referent via the same property, e.g.,

[132]In fact, I want to say that this scenario is actual, but for the point I am making all I need is that this scenario is conceptually coherent.

some neurophysiological property, but in different ways. They might, e.g., have different inferential roles.

If this scenario is coherent (conceivable), then there is no a priori reason to hold the Transparency Theses, since the Transparency Theses, each in their own way, deny the possibility of such a scenario. They all say, or at least imply, that two conceptually independent terms that refer via an essential reference fixer, could not corefer. But in fact, if this scenario is coherent, there is no reason to rule out a priori that the basic recognitional concept in our scenario is a phenomenal concept. On the account we have been considering in Chapter Two, phenomenal concepts are a special kind of recognitional concept. There is no a priori reason to rule out that a certain phenomenal concept and a certain neurophysiological concept could both refer to **pain**.

On this story the reference fixers of these terms are psychologically available, but in an important sense, they are not transparent. It is not always knowable a priori whether two terms have the same reference fixers. So, if by mode of presentation we mean whatever individuates concepts, then on this story reference fixers and modes of presentations do not coincide. Frege was wrong to suppose that there is a single entity that both individuates concepts and fixes their reference.

The above considerations show that there is no a priori reason to hold the Transparency Theses. Since the same considerations cast doubt on *Levine's Principle* as well, it is rather clear that, although his argument does not use any of the semantic machinery that the Conceivability Arguments apply, it still rests on the same basic intuitions. And those intuitions can actually be proven wrong.

The coherence of the above scenario is the key also to the actual refutation of the Conceivability Arguments. I now introduce the Zombie Refutation. This argument is a master argument: though in its original form it is directed against Jackson's argument, as I will show later, it can be easily extended to refute all the other extant (and, I hope, possible) Conceivability Arguments. What the Zombie Refutation shows is that the Conceivability Arguments are self-undermining; that is, that with the addition of some plausible further premisses we can derive a contradiction from them.

Thus, the zombies that anti-physicalists think possible in the end undermine the arguments that allege to establish their possibility. While these considerations fall short of establishing the truth of Physicalism, they go a long way towards 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.

4.2 THE MASTER ARGUMENT

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 which is exactly like our world physically, but in which no phenomenal facts obtain. 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 lead 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. Zombie-Jackson appears to give a series of lectures (as Jackson did in the actual world), arguing for the *A Priori Entailment Thesis*. What are we to make of his words?

First of all, notice that plausibly zombie-Jackson will have intentional states. When he talks, his words are not meaningless sounds. That is, it is plausible to assume that consciousness is not essential for intentionality; zombie-Jackson can have intentional states even if he lacks conscious states. Moreover, 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. On this view, zombie-Jackson's argument will be just as meaningful as Jackson's argument, and zombie-Jackson's argument, though not quite *identical* to Jackson' argument, will, in crucial respects, resemble it. It will go like this:

1*) If Physicalism is true, then for any true *T*, thoughts of the form
$K \supset T$
are conceptual truths.

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

3*) If Q^+ is a phenomenal$^+$ thought, then $K \supset Q^+$ is not a conceptual truth.

So

4*) Physicalism is false.

This argument is word by word identical to Jackson's argument; however, some of the words have different meanings in Jackson's and zombie-Jackson's mouth.[133]

[133]I marked these words with an '$^+$'. We come back to the exact nature of the difference shortly.

My argument is the following. I will argue that zombie-Jackson's argument is sound if Jackson's argument is. In particular, I will show that if a premiss of Jackson's is true, the corresponding premiss formulated by zombie-Jackson will be true as well.[134] We know, however, by assumption, that the dualist conclusion of zombie-Jackson's argument is false in the zombie-world, consequently, we know that zombie-Jackson's argument cannot be sound. But then Jackson's argument is not sound either! While this does not necessarily mean that his conclusion is false, we can conclude that the argument he uses to establish it is not effective.[135]

To run my argument, I will rely on a number of assumptions to describe the zombie-world. I am going to state these assumptions briefly right at the start; they will be discussed and defended in detail later after the argument is given. My claim is that these assumptions, taken together, are

---

[134]That is all I have to show since both arguments are clearly valid.

[135]In fact, Chalmers (1996) comes close to giving the argument himself:

> "...one might plausibly argue that a zombie does not refer to consciousness in the full sense with his word "consciousness"... But even if he does not have the full concept, there is no doubt that he judges that he has *some* property over and above his structural and functional properties - a property that he calls "consciousness"... (p. 180)

more plausible then the *A Priori Entailment Thesis*, which, if true, would lead to a dualist conclusion, a very implausible result.

Assumption 1

If the brain-states of zombie-Jackson have the same wide functional roles as Jackson's brain-states do then Jackson and zombie-Jackson share most of their intentional states except those involving phenomenal concepts.

Assumption 2

The concept that, in the zombie, corresponds to Jackson's concept *pain*, will refer to the same brain/functional state which, in the actual world, is reliably correlated with pain, and which, in the zombie-world, is also reliably correlated with the zombie's concept *pain*$^{+}$.[136] This means that whenever Jackson's thought *I am in pain* is true, zombie-Jackson's thought *I am in pain*$^{+}$ will be true as well, being about a brain state he is in.

Assumption 3

A prioricity for thoughts supervenes on the conceptual roles of the constituent concepts.

These assumptions will suffice for a reductio of Jackson's argument. As we said, zombie-Jackson, being Jackson's physical twin, offers an argument that is identical, word for word, to Jackson's argument. On *Assumptions 1* through *2*, the only difference between Jackson's and zombie-Jackson's argument is that where *Q* in Jackson's argument refers to a phenomenal fact, *Q*$^{+}$ in zombie-Jackson's argument refers to a physical fact. Premiss 1*, the *A Priori Entailment Thesis*, if true, is necessarily true, so

[136] *Pain*$^{+}$ stands for a concept of zombie-Jackson that corresponds to Jackson's concept *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.

if it was true in the actual world, it would be true in the zombie-world as well. On *Assumption 2*, we get Premiss 2*, i.e., the claim that eliminativism about phenomenal$^+$ properties is false.

Given *Assumption 3,* Premiss 3* of zombie-Jackson's argument,

(3*) If Q^+ is a phenomenal$^+$ thought then $K \supset Q^+$ is not a conceptual truth,

has as much claim to be true as Premiss 3 in Jackson's argument. While $K \supset Q^+$ has a different meaning from $K \supset Q$, it can be shown that if Premiss 3 is true then Premiss 3* is true as well. Jackson's phenomenal concepts, and zombie-Jackson's 'phenomenal$^+$ concepts have parallel conceptual roles.[137] Moreover, on *Assumption 3*, a prioricity, or conceptual necessity supervenes on the conceptual roles of the relevant concepts. That means that if $K \supset Q$ is not derivable from conceptual truths then neither is $K \supset Q^+$ derivable from conceptual truths.

[137]This is guaranteed by the epiphenomenalism with respect to qualia that is a consequence of Jackson's and Chalmers' Dualism.

On the assumptions I made, I have shown that zombie-Jackson's premisses are true if Jackson's premisses are. Premisses 1*-3* of the zombie-argument, however, together imply that Physicalism is false in the zombie-world; since this is contrary to our initial assumption, it follows that Jackson's argument must have a false premiss. The only candidate for that seems to be Premiss 1, the *A Priori Entailment Thesis*; the other premisses are extremely plausible.[138] My argument, then, by showing that the Conceivability Arguments fail, proves that Jackson's and Chalmers's principle linking conceivability and possibility is false.

I would like to consider some objections. First of all, one might object to my argument by claiming that the zombie's concept *pain*$^{+}$ *does* refer to phenomenal pain. This would be a problem for my argument. If, contrary to *Assumption 2*, zombie-Jackson's concept *pain*$^{+}$ referred to pain, then Premiss 2* would be false, since all phenomenal^{+} thoughts would be false in the zombie-world.

[138]It also follows that any argument given to support the *A Priori Entailment Thesis* must be unsound. Jackson and Chalmers used the two-dimensional account, and in particular, what we called the *A Priori Availability Thesis*, to argue for the *A Priori Entailment Thesis*. This means there is good reason to doubt the *A Priori Availability Thesis* as well, i.e., there is good reason to doubt the very semantics that is driving the Conceivability Arguments. But this is subject for another discussion.

On this view, the first intension of *pain* and *pain*$^{+}$ coincide. This is not Chalmers's view: he argues persuasively that the first intension of *pain* and *pain*$^{+}$ must be different. He points out (Chalmers 1996, p. 197)Conceivability Argument that, e.g., in spectrum inverted twins, phenomenal concepts must have different first intensions; and the zombie's concept has to be different from both of these since the zombie's concept could not distinguish between the two.

But even without Chalmers' considerations, we can see that this objection is very implausible. For zombie-Jackson, unlike for Jackson, there will be nothing *it's like* to feel pain or perceive that the sky is threatening. It is quite implausible to assume then that when Jackson thinks *that feels good*, referring to the feelings produced by a back-rub, zombie-Jackson also refers to a feeling, even though there is none in his world. Of course, I am not saying one can never have concepts that lack actual reference. The concept *unicorn* or the concept *God* seem to be a perfectly good concept. All I am claiming is that, in the particular case of phenomenal^{+} concepts, like the concept *pain*$^{+}$, the reference could not be non-physical qualia.

Pain^{+}, like *pain*, is a simple concept; its reference is not fixed via a description.[139] So if it is to refer to a non-physical property, it must be in

[139]There is a slight complication here. Georges Rey (1988) suggests that, even if the reference of *pain* is not fixed descriptively, there is a

virtue of some causal, counterfactual, or lawful relation, a recognitional ability, or some other direct, non-descriptional relation. Let's look at these possibilities one by one. Since there are no pains in the zombie-world, there could not be any causal relations between pain and the concept *pain*$^+$ either. The putative fact that the zombie's concept *pain*$^+$ refers to pain cannot be constituted by counterfactuals involving the concept *pain*$^+$ and pain either. Since Physicalism is true in the zombie-world, such counterfactual relations would have to be cashed out physicalistically.[140] But it is hard to see how this can be done with enough specificity if the counterfactuals involved are connecting physical and non-physical stuff.

The idea that the reference relation is constituted by lawful relations between the non-physical property **pain** and the zombie's concept *pain*$^+$ is not much help either. In the physicalistic zombie-world that is the subject of our discussion, there are no laws involving non-physical properties. Similarly with recognitional capacities: the zombie is not able to recognize non-

---

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. However, even if it did, we could always make up a concept that is just like *pain* except that its referent is not required, as a matter of conceptual necessity, to be non-physical. Then the argument can be run on this new concept unchanged.

[140]At least on Lewis's (1886a) account of counterfactuals.

physical phenomenal states, there being none in his world. The only remaining choice is for reference to be fixed by some other direct, non-descriptional way. Chalmers (See Chalmers 1996, p. 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, zombie-Jackson is just not acquainted with phenomenal experiences in any sense of the word.

There is another way to object to *Assumption 2,* that is, that the zombie's concept *pain*$^{+}$ refers to a brain/functional state. Instead of claiming that *pain*$^{+}$ refers to pain, one might try to argue that, even if the zombie has some intentional states, his concept *pain*$^{+}$ does not refer to anything at all. On my view, this is wrong. The natural candidate for the reference of zombie-Jackson's concept *pain*$^{+}$ is the physical or functional state that both he and Jackson occupy when the latter is having the experiences he refers to by the concept *pain*. Zombie-Jackson's thought, e.g., *that is painful*$^{+}$ attributes some brain/functional state to himself, although, of course, he does not conceive of it in this way, i.e., he does not think of this state *qua* brain-state.

Here are my reasons for this. Both Jackson and zombie-Jackson will utter the same sentences and will token the same mental representations in the same physical circumstances. Whenever Jackson uses the word (or

Mentalese expression) 'pain', zombie-Jackson will use it also; whenever Jackson's behavior is influenced by thoughts involving *pain*, zombie-Jackson's behavior is influenced by what appears to be thoughts involving *pain*+; both in Jackson and in zombie-Jackson these thoughts engage other thoughts in a way that allows for content-involving explanations. So, for example, if Jackson thinks or says he has a headache and takes an aspirin, zombie-Jackson will also token the same mental representation or utter the words 'I have a headache' and take an aspirin, etc. Finally, since certain brain/functional states of Jackson are reliably correlated with his concept *pain*, the same brain/functional states of zombie-Jackson are reliably correlated with *pain*+. I think these facts are highly suggestive of *pain*$^{+}$ referring to a brain/functional state.[141] The burden is on those who claim that these facts are not enough to establish that zombie-Jackson's *pain*$^{+}$ is a legitimate concept. The objector I am imagining here concedes that the kind of causal relations I cited in support of the claim that *pain*$^{+}$ refers to a brain/functional state are *sometimes* sufficient for

[141]In fact, Shoemaker (1998) has similarly argued that zombies will refer to a brain\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 state, since we are physically identical to our zombie-Twins.

reference.[142] She would have to give reasons why, in the particular case of phenomenal$^+$ expressions, causal relations are not sufficient for reference.

Another objection to my argument is 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.[143] The most prominent exposition of this view is due to Searle (1992, Ch. 7); he attempts to establish that consciousness is necessary for intentionality. His argument is based on considerations about the inscrutability of reference originally formulated by Quine (1960, Ch. 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. [emphasis in original] (Searle 1992, Ch. 7, p. 152)

I think this is probably wrong, and a good case can be made that zombie-Jackson does have intentional states. But even if it was true that, at

---

[142]If one denied that those causal relations are *ever* sufficient for reference, one would have to deny *Assumption 1*. I will discuss that objection shortly.

[143]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.

least potential, consciousness[144] was necessary for intentionality,[145] it would not damage my argument. My argument can be run in a way that would make the objection irrelevant.

In fact, zombie-worlds are only introduced for expository convenience. They are not essential to refute Jackson and Chalmers' argument. My argument against them only presupposes that it is conceivable to refer to a brain-state directly. I presume 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. Even on the assumption that consciousness is essential to intentionality, this will allow me to construct an analogue of the Zombie Refutation.

One way to do this is to consider a world where there are *partial zombies*. If Jackson and Chalmers is right that qualia are non-physical, then there is a world that is a physical duplicate of our world, but in which there

---

[144]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.

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

are creatures that have only some of our phenomenal experiences. These creatures will feel pleasure whenever we do, but will feel no pain at all. Since they do have consciousness, 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 they their concept *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's and Chalmers' view, possible. Imagine a world where there are creatures in many respects like us. They have the same physical and mental constitution as 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 like 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.

I think there is nothing inconceivable about this scenario. 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. In this they will be somewhat similar to actual blind-sighters; the difference is that while blind-sighters are blindly detecting some feature of their environment, yogis blindly detect some feature of their own brain. For the sake of simplicity, let us suppose that there are two different states of their brains that they can detect, state A and state B, and they use the concept *flurg* and the concept *florg* to refer directly to these states.

Yogis can formulate a variant of the Conceivability Argument. Using the *A Priori Entailment Thesis*, and the fact that truths involving the yogi-concepts, e.g., *Flurg occurred*, are not derivable a priori from the full fundamental (physical, or if Dualism is true, physical cum phenomenal) description of their world, they argue that there is a possible world exactly like theirs physically and phenomenally, but where no flurgs occur. But such a world is impossible, since, by stipulation, the concept *flurg* refers to brain state A. The yogi's argument is unsound. But among its premisses 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 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.

So, even if the objection that consciousness is essential for intentionality were sound, it would not succeed in disarming my refutation of the Conceivability Argument. But there are grounds to think that intentionality is not metaphysically dependent on consciousness. It is plausible to suppose that zombie-Jackson has thoughts, beliefs, and other intentional states, although the contents of these states need not always coincide with the contents of Jackson's states. The reason it is plausible that zombie-Jackson has intentional states is that zombie-Jackson has states that have the same wide causal functional roles as Jackson's intentional states, and such causal functional roles are plausibly sufficient for states to be intentional states. The concepts and thoughts entertained by Jackson and zombie-Jackson are identical then, except for phenomenal concepts and thoughts.[146] Zombie-Jackson will say and do exactly the same things as Jackson, and the hypothesis that he has intentional states will be just as explanatory of his

[146]If other concepts, e.g., *red*, *person*, etc., essentially involve connections to experiential concepts, as some empiricists hold, then they, too, will differ. However, their secondary intension, (i.e., their reference) may be exactly the same.

behavior as Jackson's intentional states are explanatory of Jackson's behavior. Whenever Jackson says he is thirsty and goes to the fridge for the beer, zombie-Jackson will utter the words 'I am thirsty' and reach for the beer, etc. So, there are good reasons to think that zombie-Jackson has *bona fide* intentional states.

Finally, I would like to consider an objection to *Assumption 3*, that is, the assumption that a prioricity supervenes on conceptual role. If it did not so supervene then it would be possible that sometimes we cannot tell, even in principle, 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 Premiss 3, i.e., the claim that,

for any true phenomenal thought Q, $K \supset Q$ is not a conceptual truth.

Denying *Assumption 3* would render the Conceivability Arguments extremely weak by making Premiss 3 contentious.

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 premiss that has to be given up is the *A Priori Entailment Thesis*; but on the Yogi Refutation this conclusion is inevitable.

4.3 THE EXTENSION OF THE MASTER ARGUMENT

It should be clear that the Zombie Refutation applies to the other arguments under consideration as well. We have seen how Chalmers main premiss implies Jackson's main premiss; so once we have refuted that we have refuted Chalmers' premiss as well. The BAT Argument and the Mary Argument does not require separate treatment since, as we have seen, they are inconclusive as they are stated. So we are left with Descartes', White's, Kripke's, and Levine's arguments.[147]

Zombie-Descartes is able to give a word-for-word analogue of Descartes's argument. It is quite straightforward how to argue what we have argued in the case of Jackson's argument, i.e., that the corresponding premisses in zombie-Descartes' argument are just as plausible as the premisses of Descartes's argument. The only exception is premiss 4. In premiss 4* of zombie-Descartes' argument the zombie states that he can

[147]At this point I would like to refer the reader to Appendix B, where all the arguments appearing in the dissertation are listed.

clearly and distinctly conceive of body to exist without consciousness[148] and mind to exist without extension.

But that seems problematic. What Descartes means by clear and distinct perception of a substance is to conceive of them *through their whole essence*.

The zombie, though he will *say* he has a clear and distinct perception of mind$^+$ as having (essentially) thought and consciousness$^+$, arguably does not have a clear and distinct perception of consciousness$^+$ (since he conceives of it directly without a conscious mode of presentation), and so he does not have a clear and distinct perception of mind$^+$.

However, we have independent reasons to be doubtful of Descartes' claim that one can have direct access to the whole essence of the referent of a concept just in virtue of the special features of that concept. Moreover, it will *appear* to zombie-Descartes that he has clear and distinct perceptions of mind$^+$ and body, just like it appears to Descartes that he has clear and distinct perception of mind and body. There is no reason to suppose that Descartes has better access to essences than zombie-Descartes does.

---

[148]In the original, Descartes refers to thought, which he thinks is the essence of mind in general. To increase the plausibility of the argument, I changed the premiss to state the conceivability of bodies without *consciousness*.

The Property Dualism Argument does not present any challenges to extending the Zombie Refutation to it. Premiss 3, i.e., the claim that if the same mode of presentation is associated with two (coreferring) expressions, it is knowable a priori that these expressions corefer, if true, must be necessarily true. And it is clearly true in the zombie-world that there is no physical or functional term such that it is knowable a priori that such a term corefers with *pain*$^{+}$.

Kripke's argument also straightforwardly lends itself to a Zombie Refutation. As it is prima facie possible for us that pain is not C-fibre firing, it is also prima facie possible for the zombies that pain^{+} is not C-fibre firing. Similarly, the principle that if a state of affairs is *prima facie* possible, and there is no way of explaining away this *prima facie* possibility, then this state of affairs is (metaphysically) possible, is, if true, necessarily true. And zombie-Kripke will find it just as impossible to explain away the prima facie possibility of pain^{+} not being C-fibre firing as Kripke does.

Finally, the Gap Argument. Zombie-Levine will construct an argument analogous to the Gap Argument; the Gap* Argument. Again, it is rather clear how the premisses of the Gap* Argument will have to be true, if their corresponding premisses were true in the Gap Argument. One possible objection is that it is not thickly conceivable that $K \supset Q^{+}$ is false. But if Levine cannot derive a contradiction from $-(K \supset Q)$ without the help of gapy identities,

zombie-Levine will not be able to derive a contradiction from $-(K \supset Q^{+})$ without the help of gapy identities either.

Suppose that zombie-Levine's concept *pain*$^{+}$ refer to C-fibre firing, so that $-K \supset$ *x has pain*$^{+}$ is true. This will be a gapy identity for zombie-Levine and any derivation of a contradiction from $-K \supset$ *x has pain*$^{+}$ must include a gapy identity.

I conclude that the all the Conceivability Arguments, as well as the Gap Argument, fail.

4.4 "Explaining Away" the Mind-Body Problem

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 the 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 premisses are plausible, except the premiss 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, pp. 134-5).

With the Zombie Refutation and its companion arguments, we can actually do better. The argument actually shows where the anti-physicalist 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*:

(E) For any true statement T
$\Box(K \supset T)$.

The explanation that Jackson puts forward of why E holds is that all instances of E are conceptual truths. He thought that the reduction of higher level concepts to lower level concepts has to be *perspicuous.*[149] 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.

The other problem is this. Many of us are convinced (partly by the Conceivability Arguments) that there is something special about phenomenal

[149]See Chapter One for a detailed discussion of physicalism and reductionism.

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.[150] 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 this 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 the account of phenomenal concepts we adopted in Chapter Two, it is *to be expected* for a physicalist that there will be an explanatory gap, and that zombies are conceivable.

[150]If you prefer the Gap Argument, think about it in terms of thin and thick conceivability.

On this account, now supplemented with physicalism, we get the following picture. 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, micro physical) 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 which 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 micro physics, each of which pick out their 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 Transparency Theses.[151]

But what about the persistent intuition (among lay people and philosophers alike) that, despite of 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

[151]A similar point is made by Scott Sturgeon (1994).

Transparency Theses.[152] It is a very powerful intuition that, if we have two concepts, both of which refer via an essential reference fixer, then we *must* be able to tell if they corefer. After all, we have an insight into the nature of their referent through their reference fixers; so how could we be wrong about our judgement (as it is in the phenomenal/neurophysiological case) that they do not corefer? But we have seen in section 4.1 that this intuition is misplaced.

[152]Papineau (1993a) argues along similar lines. He gives an account which he claims explains why the peculiarities of our phenomenal concepts inevitably give rise to the illusion of Dualism.

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 (non-phenomenal) physical states. Given my refutation of the Conceivability Arguments, I cannot claim, merely on the basis of their conceivability, that they are possible.[153] But I see no reason why they could not be; after all, blind sight experiments show that actual people possess concepts that are in many ways similar to the yogi concepts. [154] Yogis can make 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 shouldn't they be, given our claim that a belief in the Transparency Theses is enough to explain anti-physicalist intuitions? Yogis are just as attracted by the Transparency Theses as ordinary humans are. The answer to this objection is that, when all is said and done, the yogi can be attracted by the Transparency Theses, and still not drawn to Dualism, just

---

[153]Though mere conceivability was enough to make the point against the Conceivability Arguments.

[154]The main difference is, ignoring the fact that blind-sighters are reluctant to employ the recognitional abilities that constitute their blind-sight concepts without prompting, is that in the case of blind-sight we have direct recognitional concepts of external objects and properties, in the case of yogis, we have direct recognitional concepts of internal states (brain-states).

on the basis of her special conceptual repertoire. The difference between us vis a vis phenomenal concepts, and the yogi vis a vis the yogi concepts is that, as opposed to us, the yogi does not have a temptation to think that she has direct insight into the nature of what her concepts *flurg* and *florg* refer to. She does not have a real handle on the concept; the reference fixer of her concept might be an essential property of the referent, but she does not have an access to it, the way we have access to the phenomenal reference fixers of our phenomenal concepts.

To conclude, I think we can sum up our project thus: if we have not completely untied the “World-knot” of the Mind-Body problem, we have loosened it a bit.

APPENDIX A: IMPORTANT DEFINITIONS

Conceivability

(Con) A statement S is conceivable, if it is logically consistent with the totality of conceptual truths, i.e., if -S is not a conceptual truth.

Physicalism

P Among worlds where no property alien to the actual world is instantiated, any two that are exactly alike with respect to their complete descriptions (including specification of the fundamental laws) in the language of the ideal fundamental physical theory are duplicates simpliciter.

Entailment Thesis

(E) For any true statement T,
$\Box(K^*\&F\supset T)$,

where K* is the full fundamental description of the world, F is the claim that K* *is* the full fundamental description of the world, and T is any truth.

Physicalist Entailment Thesis

(E) For any true statement T,
$\Box(K\&F\supset T)$,

where K is the full physical description of the world, F is the claim the K *is* the full fundamental description of the world, and T is any truth.

Transparency Theses

Cartesian Transparency Thesis

(CTT) When you clearly and distinctly perceive of substance A and substance B, that is, you conceive of them *through their whole essence*, and still do not see that they are *the same substance*, then they must be different.

White's Transparency Thesis

(WTT) If the same mode of presentation is associated with two (coreferring) concepts, it is knowable a priori that these concepts corefer.

Kripkean Transparency Thesis

(KTT) it is impossible to refer to the same kind of state through rigid designators with essential reference fixers that have no conceptual connection at all.

A Priori Entailment Thesis:

(APET) if (E) is true, for any true T, statements of the form
$K \supset T$
are conceptual truths.

Conceivability-Possibility Thesis

(C-P) if Physicalism is true then, for any statement S, if 'K&S' is conceivable then it is possible.

Necessity-Contingency Thesis

(NCT) the primary intension of a necessary a posteriori thought must be contingent.

Levine's Principle

(LP) Thickly conceivable situations are metaphysically possible.

APPENDIX B: THE CONCEIVABILITY ARGUMENTS

DESCARTES' ARGUMENT FOR THE REAL DISTINCTION BETWEEN MIND AND BODY

1) Whatever I can clearly and distinctly understand can be brought about by God.

2) If α belongs to the essence of A and β belongs to the essence of B, and I can clearly and distinctly understand B to exist without α and A to exist without β, then I can clearly and distinctly understand A to exist without B and B to exist without A.

3) Thought belongs to the essence of mind, extension belongs to the essence of body.

4) I can clearly and distinctly understand body to exist without thought and mind to exist without extension.

5) By (1), (2), (3) and (4), mind can exist apart from body.

6) If A can exist apart from B, and vice versa, A is really distinct from B.

7) Hence, by (5) and (6), mind is really distinct from body.

ARGUMENT FROM DOUBT

1) I am certain that my mind exists.

2) I am not certain that my body exists.

3) My mind is diverse from my body.

NAGEL'S BAT-ARGUMENT

1) There are no physical facts that do not consist in the truth of propositions expressible in a human language.

2) There are experiences that we cannot conceive of.

3) Having an experience is a fact.

4) If there are facts that we cannot conceive of then there are facts that do not consist in the truth of propositions expressible in a human language.

Lemma (by 2, 3, and 4):

5) There are facts that do not consist in the truth of propositions expressible in a human language.

So, by 1and 5,

6) There are non-physical facts.

JACKSON'S MARY-ARGUMENT

1) Mary knows all the physical facts.

2) Mary doesn't know what it is like to experience red.

3) What it is like to experience red is a fact.

Lemma:

4) There is a fact that Mary doesn't know.

Conclusion:

5) There are non-physical facts.

PROPERTY DUALISM ARGUMENT

1) An expression refers to an entity via a mode of presentation. The mode of presentation provides the route by which the entity is picked out by the expression.

2) Modes of presentation are properties of the referent.

3) If the same mode of presentation is associated with two (coreferring) expressions, it is knowable a priori that these expressions corefer.

4) The concept *pain* is a referring expression.

5) Physical or functional concepts have as their mode of presentation physical or functional properties.

6) There is no physical or functional term such that it is knowable a priori that such a term corefers with *pain*.

Lemma:

7) No physical or functional property of pain could provide the route by which pain is picked out by the expression *pain*.

8) Properties are either physical, or functional, or irreducibly mental.

Conclusion:

9) Pain has at least one property that is irreducibly mental.

KRIPKE'S ARGUMENT FOR DUALISM

1) It is *prima facie* possible that C-fibre firing is not pain.

2) If a state of affairs is *prima facie* possible, and there is no way of explaining away this *prima facie* possibility, then this state of affairs is (metaphysically) possible.

3) There is no way of explaining away the *prima facie* possibility of C-fibre firing not being pain.

4) By 1, 2 and 3, it is (metaphysically) possible that C-fibre firing is not pain.

5) But if it is (metaphysically) possible that pain is not C-fibre firing, then, since both "pain" and "C-fibre firing" are rigid designators, pain is not C-fibre firing.

6) By 4 and 5, pain is not C-fibre firing.

JACKSON'S ARGUMENT

1) If Physicalism is true, then for any true T, statements of the form

$K \supset T$

are conceptual truths.

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

3) If Q is a phenomenal statement , then '$K \supset Q$' is not a conceptual truth.

So

4) Physicalism is false.

CHALMERS' ARGUMENT

1') If Physicalism is true, then for any true statement T

(E) $K \supset T$

is necessarily true,

where K is the complete physical description of the world.

2') If '$K \supset P$' is not a conceptual truth then the primary intension of '$K \supset P$' expresses a contingent proposition.

3') If Q is a statement expressing a phenomenal fact in the language of psychology, then '$K \supset Q$' is not a conceptual truth.

So (from 2' and 3')

4') The primary intension associated with '$K \supset Q$' expresses a contingent proposition.

5')The primary and secondary intension of micro-physical statements express the same proposition.

So (from 4' and 5')

6') '$K \supset Q^{-}$', where Q^{-} is the primary intension of the statement Q, expresses a contingent proposition.

So (from 6')

7') '$K \supset Q^{-}$' is not necessarily true.

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

9') The primary proposition of a true statement is true as well.

So (from 1', 7', 8' and 9')

10') Physicalism is false.

THE GAP ARGUMENT

1) If physicalism is true, then for any true statement T, $\Box(K \supset T)$.

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

3) It is thickly conceivably that $K \supset Q$ is false.

4) Thickly conceivably situations are metaphysically possible.

5) Physicalism is false.

BIBLIOGRAPHY

David Armstrong (1968): *A Materialist Theory of the Mind*, Routledge and Kegan Paul.

David Armstrong (1978): *A Theory of Universals*, Cambridge UP.

David Armstrong (1983): *What is a Law of Nature?*, Cambridge, Cambridge University Press.

Bigelow, J & Pargetter, R (1990): "Acquaintance with Qualia", *Theoria* 56:129-47.

Ned Block and Jerry Fodor (1972): "What Psychological States Are Not", in *Philosophical Review* 81:159-81.

Ned Block, (ed.) (1980a): Readings in Philosophy of Psychology, Harvard University Press.

Ned Block (1980b): "Are Absent Qualia Impossible?", *The Philosophical Review* 89: 257-274

Ned Block (1986): “Advertisement for a Semantics for Psychology”, in P. French, T. Euhling an H. Wettstein (eds), *Studies in the Philosophy of Mind*, vol 10 of *Midwest Studies in Philosophy*, pp. 615-78.

Ned Block (1990): "Inverted Earth", *Philosophical Perspectives,* 4, Action Theory and Philosophy of Mind.

Ned Block (1994a): "On a Confusion about a Function of Consciousness", *Behavioral and Brain Sciences* 17:2

Ned Block (1994b): "Consciousness" in *A Companion to the Philosophy of Mind,* (ed. S. Guttenplan), Basil Blackwell, Cambridge, MA.

Ned Block and Robert Stalnaker (1997): "Conceptual analysis and the explanatory gap", (forthcoming).

Paul Boghossian (1996): "Analyticity reconsidered", *Noûs* 30:3, pp. 360-391.

Tyler Burge (1988): “Individualism and Self-Knowledge”, in *Journal of Philosophy*, 85: 649-63.

Rudolf Carnap (1955): *Meaning and Necessity: A Study in Semantics and Modal Logic*, 2nd ed., Chicago: University of Chicago Press.

David Chalmers (1996): *The Conscious Mind*. Oxford UP.

Roderick Chisholm (1957): *Perceiving*. Ithaca, N.Y.: Cornell University Press.

Noam Chomsky (1975): *Reflections on Language*, Pantheon Books: New York.

Paul Churchland (1981): "Eliminative materialism and propositional attitudes", *The Journal of Philosophy*, 78:67-90.

John Cottingham; Robert Stoothoff; Dugald Murdoch (1984): *The Philosophical Writings of Descartes*, Cambridge UP.

Tim Crane and D.H.Mellor (1990): “There is No Question of Physicalism.”, in *Mind* 99, pp. 185-206.

Martin Davies and Lloyd Humberstone (1980), "Two Notions of Necessity", *Philosophical Studies* 38:1-30.

Daniel Dennett (1991a): *Consciousness explained,* Little, Brown and Co, Boston.

Daniel Dennett (1991b): ""Epiphenomenal" Qualia?:", in *Consciousness Explained* (Little-Brown), pp. 398-406.

Fred Dretske (1977): “Laws of Nature”, in *Philosophy of Science*, 44, pp. 246-68.

Fred Dretske (1995): *Naturalizing the Mind*, Cambridge, MA: MIT Press.

Gareth Evans (1979): "Reference and contingency", *The Monist* 62:161-89.

Herbert Feigl (1967): *The `Mental' and the `Physical',* Minneapolis: University of Minnesota.

Jerry Fodor (1975): *The Language of Thought, New* York: Crowell.

Jerry Fodor (1987): *Psychosemantics*, Cambridge, MA, MIT Press.

Jerry Fodor (1990): *A Theory of Content and Other Essays*, MIT Press/Bradford Books.

Jerry Fodor (1994): *The Elm and the Expert: Mentalese and Its Semantics*, Cambridge, Mass.: MIT Press

Jerry Fodor (1997): *Concepts*, Locke Lectures.

Peter Geach (1957): *Mental Acts*, London: Routledge and Kegan Paul.

Samuel Guttenplan (ed.) (1994): *A Companion to the Philosophy of Mind*, Blackwell.

Gilbert Harman (1990):"The intrinsic quality of experience", *Phil Perspectives* 4, Action Theory and Philosophy of Mind, pp. 31-52.

Gilbert Harman (1993): "Can Science Understand the Mind?" in Harman (ed.), *Conceptions of the Mind,* Essays in Honor of George A. Miller, Lawrence Erlbaum Associates, Publishers, Hillsdale, New Jersey, 1993, pp. 111-21.

Terry Horgan (1982): "Supervenience and microphysics", *Pacific Philosophical Quarterly* 63: 29-43.

Frank Jackson (1980): "A note on Physicalism and heat", *Australasian Journal of Philosophy,* Vol 58, No. 1.

Frank Jackson (1982): "Epiphenomenal Qualia", *Philosophical Quarterly* 32:127-136. Reprinted in (W. Lycan, ed.): *Mind and Cognition* (Blackwell, 1990).

Frank Jackson (1986): "What Mary did not Know", *Journal of Philosophy* 83, pp. 291-5.

Frank Jackson (1993): "Armchair Metaphysics", in (M. Michael and J. O'Leary-Hawthorne, eds.): *Philosophy in Mind*, Kluwer Academic Publishers.

Frank Jackson (1995): "Postscript to "What Mary did not know"". In P. K. Moser and J. D. Trout, eds., *Contemporary Materialism*. London: Routledge.

Frank Jackson (1982): "Epiphenomenal Qualia", *Philosophical Quarterly* 32:127-136. Reprinted in (W. Lycan, ed.): *Mind and Cognition* (Blackwell, 1990).

Frank Jackson (1993): "Armchair Metaphysics", in (M. Michael and J. O'Leary-Hawthorne, ed.): *Philosophy in Mind*, Kluwer Academic Publishers.

Frank Jackson (1995): Locke Lectures, April 1995.

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

Henry Jacoby (1989): "Empirical Functionalism and Conceivability Arguments", *Philosophical Psychology*, Vol. 2, No. 3, pp. 271-282.

David Kaplan (1978): "On the logic of demonstratives." *Journal of Philosophical Logic* 8:81-98.

David Kaplan (1979): "Dthat". In (P. Cole, ed) *Syntax and Semantics*. New York: Academic Press.

David Kaplan (1989): "Demonstratives", in J. Almog, J. Perry, and H. Wettstein, eds., *Themes from Kaplan*. New York: Oxford University Press.

Jagwon Kim (1990): “Supervenience as a Philosophical Concept”, reprinted in *Supervenience and Mind*, Cambridge, Cambridge University Press, 1993.

Jagwon Kim (1993): *Supervenience and Mind*, Cambridge UP.

Saul Kripke (1972): *Naming and Necessity* (Harvard UP).

Janet Levin (1986): "Could Love Be Like a Heatwave? Physicalism and the Subjective Character of Experience", *Phil Studies,* Vol. 49, No.2, pp. 245-61.

Joseph Levine (1983): "Materialism and Qualia: The Explanatory Gap", *Pacific Philosophical Quarterly*, 64, 354-361.

Joseph Levine (1993): “On Leaving Out What It’s Like”, in Davies, M. And Humphreys, G., ed., *Consciousness: Psychological and Philosophical Essays.* Oxford: Blackwell, 121-136.

Joseph Levine (1998): *Conceivability and the metaphysics of mind*, Manuscript.

David Lewis (1966): "An argument for the identity theory. *Journal of Philosophy* 63:17-25.

David Lewis (1979): "Attitudes *de dicto* and *de se*", *Philosophical Review*, 88: 513-45.

David Lewis (1983): "New work for a theory of universals", *Australasian Journal of Philosophy* 61:343-77.

David Lewis (1986a): *On the Plurality of Worlds*, Basil Blackwell.

David Lewis (1986b): *Collected Papers*, Vol. 2, Oxford University Press, New York.

David Lewis (1988): "What Experience Teaches", in *Proceedings of the Russellian Society*, (ed. J. Copley-Coltheart), University of Sidney. Reprinted in *Mind and Cognition*, ed. W. Lycan: 499-518. Oxford: Blackwell, 1990.

David Lewis (1994): "Reduction of mind", in (S. Guttenplan, ed.) *A Companion to the Philosophy of Mind*, Blackwell.

Brian Loar (1988): "Social content and psychological content", in *Contents of Thought*, ed. R.H. Grimm and D.D. Merrill, Tucson: University of Arizona Press.

Brian Loar (1990): "Phenomenal states", *Philosophical Perspectives* 4, Action Theory and Philosophy of Mind, pp. 81-108.

Brian Loar (1997): “Phenomenal states”, in *The Nature of Consciousness*, (Block, Flanagan, Güzeldere, eds.), MIT Press (revised version of Loar (1990)).

Barry Loewer and Ernie LePore (1985): “Solipsistic Semantics”, *Midwest Studies in Philosophy*.

Barry Loewer (1995): "An Argument for Strong Supervenience", in: *Supervenience*, (ed. Elias E. Savellos, Ümit D. Yalcin), Cambridge UP.

Barry Loewer (1997): “Humean Supervenience”, in *Philosophical Topics*.

W.G. Lycan (1996): *Consciousness and Experience,* MIT Press.

David Marr (1982): *Vision*, San Francisco: D.H. Freeman and Co.

Tim Maudlin (1989): "Computation and consciousness", *The Journal of Philosophy* 86:407-432.

John Stuart Mill (1843): *A System of Logic*, New York, Harper and Brothers, 1893.

Colin McGinn (1977): "Anomalous monism and Kripke's Cartesian intuitions", *Analysis* 37, no. 2, pp. 78-80.

Colin McGinn (1982): *The Character of Mind*, Oxford: Oxford University Press.

Colin McGinn (1987): "Critical notice: *The View from Nowhere* by Thomas Nagel" in *Mind* 96:263-272.

Thomas Nagel (1974): "What is it like to be a bat?", *Philosophical Review* 4:435-50.

Thomas Nagel (1979): *Mortal Questions,* Cambridge University Press.

Thomas Nagel (1986): *The View From Nowhere,* Oxford UP.

Thomas Nagel (1993): "What is the mind-body problem?", in *Experimental and theoretical studies of consciousness.* Wiley, Chichester (Ciba Foundation Symposium 174) pp 1-13.

Karen Neander (1992): "Sensational Knowledge", (Manuscript).

Laurence Nemirow (1990): "Physicalism and the Cognitive Role of Acquaintance", in *Mind and Cognition* (ed. W. Lycan), Basil Blackwell, Cambridge, MA.

David Papineau (1993a): "Physicalism, Consciousness, and the Antipathetic Fallacy", *Australasian Journal of Philosophy* 71:169-83.

David Papineau (1993b): *Philosophical Naturalism*, Blackwell.

Christopher Peacocke (1992): *A Study of Concepts*, Cambridge, MA: MIT Press.

John Perry (1979): "The problem of the essential indexical", *Nous* 13:3-21

Hilary Putnam (1967): "The Nature of Mental States". In W.H. Capitan and D.D. Merrill, eds., *Art, Mind, and Religion*. Pittsburgh: University of Pittsburgh Press, pp. 37-48.

Hilary Putnam (1975): "The meaning of meaning", in K. Gunderson, ed., *Language, Mind, and Knowledge*. Minneapolis: University of Minnesota Press.

W.V. Quine (1951): "Two dogmas of empiricism", *Philosophical Review* 60:20-43.

W.O. Quine (1960): *Word and Object*, MIT Press.

W.V. Quine (1969): "Propositional objects". In *Ontological Relativity and Other Essays*. New York: Columbia University Press.

Georges Rey (1983): "A Reason for Doubting the Existence of Consciousness", in Davidson, R., Schwartz, G. and Shapiro, D., eds., *Consciousness and Self-Regulation,* vol. III, New York: Plenum, pp. 1-39.

Georges Rey (1988): “A Question about Consciousness”, in: *Perspectives on Mind*, ed. H. Otto and J. Tueidio, Kluwer Academic Publishers.

Robinson, H (1993): "The Anti-materialist Strategy and the `Knowledge Argument'", in (H. Robinson, ed.) *Objections to Physicalism* (Oxford UP).

David Rosenthal (1990): "A Theory of Consciousness". *Report No. 40,* Center for Interdisciplinary Research (ZIF), Research Group on Mind and Brain, University of Bielefeld.

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

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

Sydney Shoemaker (1975): "Functionalism and Qualia", *Philosophical Studies* 27:291-315.

Sydney Shoemaker (1979): “Identity, Properties and Causality”, in *Identity, Cause and Mind*, Cambridge University Press, 1984, pp. 234-60.

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

Sydney Shoemaker (1982): "The inverted spectrum", *Journal of Philosophy* 79, pp. 357-381.

Sydney Shoemaker (1998): Commentary in “Symposium on Chalmers’ *The Conscious Mind*”, forthcoming in *Philosophy and Phenomenological Research*.

B.F. Skinner (1953): *Science and Human Behavior*, New York: Macmillan.

J. J. C. Smart (1959): "Sensations and Brain Processes." *Philosophical Review* 68:141-156.

Stalnaker, R (1978): "Assertion". (P. Cole, ed.) *Syntax and Semantics: Pragmatics, Vol. 9.* Academic Press, 1978.

Scott Sturgeon (1994): “The Epistemic View of Subjectivity”, *Journal of Philosophy*, vol. XCI, no 5: 221-236.

Michael Tye (1995): *Ten Problems of Consciousness: A Representational Theory of the Phenomenal Mind.* MIT Press.

Michael Watkins (1989): "The Knowledge Argument Against the Knowledge Argument", *Analysis* 49:158-60.

Stephen White (1986): "Curse of the Qualia", *Synthese* 68, pp. 333-68.

Eugene Wigner (1967): *Symmetries and Reflections*, MIT Press, Cambridge.

Margaret Wilson (1978): *Descartes*, Routledge and Kegan Paul.

Gene Witmer (1997): *Demanding Physicalism*, Doctoral Dissertation, Rutgers University.

Ludwig Wittgenstein (1953): *Philosophical Investigations*, Basil Blackwell.

Stephen Yablo (1993): "Is conceivability a guide to possibility?", *Philosophy and Phenomenological Research*, Vol. LIII, No. 1.