CONSCIOUSNESS, CONCEIVABILITY AND CONCEPTS

Encarnación Díaz León

Thesis submitted for the degree of Doctor in Philosophy

Department of Philosophy
University of Sheffield

August 2007

Summary

The main aim of this thesis is to defend the metaphysical doctrine of *physicalism* (which claims that everything is physical) from a very powerful class of anti-physicalist arguments, namely, the so-called *conceivability arguments*. These arguments are concerned with *phenomenal consciousness*, that is, the aspect of mentality having to do with what it is like to undergo mental states. Conceivability arguments purport to show that phenomenal consciousness is not physical, and therefore, physicalism fails since there is at least one phenomenon that is not physical.

Physicalism is usually characterised in terms of a *supervenience* thesis: physicalism is true if and only if all facts (including phenomenal facts) supervene on physical facts. Conceivability arguments try to show that the phenomenal does not supervene on the physical, that is, that there is at least one possible world that is physically identical to ours but where phenomenal facts differ.

Conceivability arguments start by stating that we can *conceive* of a possible world physically identical to the actual world, but where phenomenal facts are different (for instance, zombie worlds). From this, they infer that such a world is *possible*. That is, they think that the conceivability of such a world entails its possibility. And as just noted, if such worlds are possible, physicalism fails.

My goal is to argue that the *conceivability-possibility link* posited by advocates of conceivability arguments (such as Chalmers, Jackson, Kripke, White and others) is not correct. I examine two different strategies against such inference from conceivability to possibility, namely, the so-called *exceptionalist* and *non-exceptionalist strategies*, and I develop and defend both of them. One of my objectives is to show that the best over-all strategy is a combination of both exceptionalism and non-exceptionalism, which, I argue, is perfectly coherent.

A mis hermanas y mi padre, y en memoria de mi madre

(To my sisters and my father, and in memory of my mother)

Acknowledgements

Writing a PhD is no small task. Many people have helped and supported me through these years, and I wish to thank them for their time and generosity.

First of all, I am very grateful to my supervisors, Stephen Laurence and Rosanna Keefe, for their support and encouragement and for their extremely helpful comments and advice. They have both patiently gone through (far too) many earlier drafts of this material, and have always helped me to get the best out of my ideas and to explain them as clearly as possible. Special thanks are due to my primary supervisor, Stephen Laurence, who encouraged me to come to Sheffield to do a PhD in the first place, and who has been extremely supportive and generous with his time at all stages of this PhD. Thanks also to Dominic Gregory, who took over as my secondary supervisor for one semester, for his sharp and helpful comments on my work and for reading some of my papers afterwards.

I am also grateful to the *Arts and Humanities Research Council*, the University of Sheffield and the *Royal Institute of Philosophy*, for their generous financial support.

The Department of Philosophy at the University of Sheffield has proved to be a very congenial place to work. I have had the chance to present many of the ideas in this work at the Graduate Research Seminars, where I have always received invaluable feedback. I am particularly grateful to the following for very helpful comments and discussion: Andre Abath, Giles Banning-Lover, John Divers, Simon Fitzpatrick, Fabian Freyenhagen, Jules Holroyd, Rob Hopkins, Gerry Hough, Rosanna Keefe, David Liggins, Anna Mahtani, Joe Morrison, Julien Murzi, Tina Mussgnug-Barrett, Davide Rizza, Doug Ryan, Jennifer Saul, Andrew Thomas, Ian White and Richard Woodward. (I apologize to those that I have forgotten to mention). I am especially grateful to Laura Beeby, Simon Fitzpatrick, Andrew Howat, David Liggins and Lina Papadaki, for proof-reading the thesis. Extra thanks are due to David Liggins for so many enjoyable and fruitful discussions on metaphysics, meta-metaphysics and beyond.

I also had the chance to teach a final-year undergraduate course on the topic of this thesis, which was an excellent way of trying out some of my ideas. I am indebted

to my Department for this opportunity, and to my students for their interest, enthusiasm, and many interesting discussions.

I have also presented some of this material in several workshops and conferences at different venues, including: Barcelona, Girona, Leeds, Lisbon, Murcia, Oviedo, San Raffaele (Milan), Southampton, UAB and Valencia. I am very grateful to the audiences in all those occasions for very useful comments and discussion. I can remember especially helpful comments from: Alex Buckley, Oscar Cabaco, Andy Clark, David Chalmers, Jose Díez, Angel García Rodríguez, Manuel García-Carpintero, Philip Goff, Sandford Goldberg, Harris Hatziioannou, Luca Incurvati, Dan López de Sa, Manolo Martínez, Chris Peacocke, Manuel Pérez Otero, David Pineda, Karol Polcyn, Jose Luis Prades, Luis Robledo, Sonia Roca, Pablo Rychter, Jordi Valor and Elia Zardini. Extra thanks are due to Saray Ayala, Marta Moreno, Luis Robledo and Sonia Roca, who wrote replies to my talks in Valencia, Girona, UAB and Barcelona, respectively. I am especially indebted to David Chalmers, who read some of this material and provided very helpful and illuminating comments.

Special mention is due to Dan López de Sa, who has read most of this material in some form or another, and who has provided extremely helpful feedback at all stages of this project. He is the one who first introduced me to analytic philosophy, and I warmly thank him for that, and for his endless encouragement, support, advice, patience, generosity, fierce criticism, sense of humour and friendship. Extra thanks also for suggesting the title.

I would also like to thank all the friends, many of them postgraduate and/or visiting students at the Department of Philosophy, who have made Sheffield feel like home during all these years, and very especially, Lina Papadaki, who has been here from the beginning to the end, and who, I hope, will continue to be.

Finally, I wish to thank my friends and family back at home, for being so encouraging, supportive and faithful. I wish to dedicate this work to my sisters and father, for their generosity and dedication, and especially to the memory of my mother, who could not see this work finished, and to whom I owe so much.

Contents

1. What is Physicalism?

1.1. Introduction

Physicalism is the claim that everything is physical, that is, everything that exists in the actual universe is physical; there is nothing over and above the physical. In this thesis I will discuss some important arguments against physicalism, but before we can properly assess the truth or falsity of physicalism, it will be useful to have a clear understanding of what physicalism amounts to. In particular, in this chapter I aim to offer an answer to the question: what does it *mean* to say that everything is physical? Once we have a satisfactory answer to that question, we can proceed to the question of whether physicalism is true or not, or more specifically, whether certain arguments against physicalism are successful or not.

In this chapter we will evaluate some different characterizations of physicalism that have been discussed in the literature, and we will focus on a very influential characterization, which we will label *Physicalism*. As we will see, the central aim for the purposes of this chapter is to show that the intuitive claim of physicalism (that is, the claim that everything is physical) *entails* the truth of Physicalism. It is not so clear, though, whether Physicalism itself can entail the intuitive claim of physicalism. However, this will turn out to be irrelevant for our purposes here, since the anti-physicalist arguments that we will explore in detail in the following chapters will be directed against the specific claim of Physicalism. If these arguments succeed in showing that Physicalism is indeed false, then it clearly follows that the intuitive claim of physicalism fails too, since the latter entails the former. In this chapter, I will explain what Physicalism says and why it is entailed by the claim that there is nothing over and above the physical. In the following chapters, we will explore certain arguments against Physicalism, and I will argue that they ultimately fail. This will contribute to the defence of physicalism, as intuitively conceived.

The question of what it means to say that everything is physical can be naturally divided into two questions: the completeness question and the condition question.[1] The *completeness question* is the following:

[1] Here I follow the excellent discussion in Stoljar (2005a). See also Dowell (2006) for a similar distinction.

- What does it mean to say that *everything* is physical?

That is, the completeness question asks what it is *for everything* to satisfy one condition or another, such as being physical, or whatever. What does it mean to say that everything in the universe satisfies a certain condition? This question not only arises for the claim of physicalism but also for other metaphysical claims such as Berkeley's idealism, which says that everything is mental. Likewise, this claim gives rise to the question: what does it mean to say that *everything* satisfies the condition of being mental? We will assess different answers to the completeness question, which try to clarify what it means for everything in the universe to satisfy a given condition.

The *condition question* asks the following:

- What does it mean to say that everything is *physical*?

That is, the condition question assumes that we understand what it means for everything to satisfy one condition or another, and asks instead what it means for something (everything, in this case) to satisfy the condition of *being physical*. What is this condition that according to physicalism, everything satisfies? What condition should something satisfy in order to be physical? Again, a parallel question could be raised with respect to the claim of idealism: what is the condition of *being mental* that according to idealism, everything satisfies?

We will first examine the completeness question, and then turn to the condition question. In order to examine the completeness question, it is useful to bracket the condition question, that is, we will assume that we understand what it is for something to be physical. A good way to proceed is to posit an intuitive understanding of what physical is, so as to examine the completeness question. Later, we will explore some problems of such an intuitive understanding, and we will see how to refine it.

1.2. The Completeness Question: Identity and Supervenience

The most popular answer to the completeness question in contemporary philosophy is based on the notion of *supervenience*. I will present an intuitive characterization of this notion, by comparing it with another, simpler answer to the completeness

question. I will then introduce some of the main attempts to cash out the intuitive idea of supervenience more precisely, and I will assess them with respect to the purposes of defining physicalism.

1.2.1. Type-physicalism and token-physicalism

The simplest answer to the completeness question, that is, the question of what it is for *everything* to be physical, is perhaps the following:

> (I): Everything is physical if and only if all entities in the universe are physical entities.

From this answer, we can obtain our first definition of physicalism, as follows:

> (*Physicalism I*): All entities in the universe are physical entities.

Of course, this answer to the completeness question presupposes that we understand what physical entities are. As I said, we will examine this question in more detail when we discuss the condition question, but for the time being, it will suffice to give this rough characterization: we can say that physical entities are those posited by physics (or the physical sciences).[2]

The main problem with (I) as an answer to the completeness question is the following. We usually take for granted the existence of many things in the universe: tables, chairs, storms, rivers, the colour blue, pain, and so on. It seems obvious that these entities are not the kind of entities that are posited by physics: these are rather entities such as atoms, protons, electrons, etc., and properties such as spin, charge, mass, and so on. That is, the entities and properties that are invoked in physics are rather more microscopic, so to speak, than the entities we are concerned with in everyday discourse. If physicalism were understood as the claim Physicalism I (i.e. that all entities that exist in the universe are posited by physics) then we would have a dilemma: either we reject physicalism (because there exist entities that are not posited

[2] In section 1.3 we will see some problems concerning this characterization, and some attempts to improve it.

by physics, such as chairs), or we reject the existence of all the entities and properties that are not posited by physics, which would be very counterintuitive.[3]

A good solution here is the following: what we have to do in order to answer the question of what it means for *everything* to be physical is to find a relation *R* such that if something bears *R* to the entities and properties posited by physical science, then it is physical in an intuitive sense. That is, we are looking for a way of expanding our ontology beyond the entities and properties posited by physics, but which still respects the physicalist idea that there is nothing over and above the physical. Thus, we can reformulate the completeness question as the search for this relation *R*, and we can then say that everything is physical if and only if everything bears *R* to those entities posited by physics. In this section, we will examine two possible answers: one is *identity*, the other is *supervenience*.

The first option goes as follows: we could accept that there exist many entities in the actual world that are not explicitly invoked in physical theory, but we could hold that they can be *identified* with entities that are explicitly invoked. According to this idea, we could answer the completeness question in the following way:

> (I*): Everything is physical if and only if all entities in the universe are identical to physical entities.

There are two ways of understanding this general idea, which in turn give rise to two different notions of physicalism. According to the first one, so-called *token-physicalism*, every particular entity or token in the actual world is identical to some physical token. According to the second one, so-called *type-physicalism*, every type or property that is instantiated in the actual world is identical to some physical type.

Can we capture the intuitive idea of physicalism in terms of either of these two notions? Let's examine type-physicalism first. Does physicalism entail that all properties are identical to physical properties? That is, does the intuitive claim of physicalism, namely, the idea that there is nothing over and above the physical, entail type-physicalism? It seems reasonable to say that physicalism does not really entail this. One reason for this is the following: there are many properties that seem to be *multiply realizable*. That is, the same property can be realized by different *physical*

[3] See Melnyk (2003: 67-8) for a very good exposition of this problem.

properties. For instance, the property of being a chair seems to be multiply realizable in this sense. Chairs can be made out of different materials: wood, metal, plastic, etc. That is, there are chairs that instantiate very different physical properties, but all of them instantiate the property of being a chair, so being a chair is multiply realizable. Therefore, there seems to be no specific physical property that could be identified with the property of being a chair. However, this seems to pose no problem for the idea that there is nothing *over and above* the physical. Physicalism, as intuitively conceived, can still be true, even if there are many properties which cannot be identified with any property posited by physics.

Therefore, physicalism is not committed to type-physicalism. What about token-physicalism? Can we say that the idea that there is nothing over and above the physical is committed to the claim that every token in the actual world is identical to some physical token, posited by physics? One problem with this view is the following. Let's consider the example of a statue and the lump of clay it is made of. Can we say that the statue is *identical* to such a lump of clay (which we can assume is a physical token)? It is not clear whether this is so: some philosophers claim that the relation between the statue and the physical stuff it is made of is not identity but something else (e.g. the statue is not identical to the lump of clay but rather *constituted* by it). This is a complicated issue which we do not have to assess here. The main idea for our purposes is that even if it turned out that the statue is constituted by the lump of clay rather than being identical to it, this does not seem to pose a problem for the claim that everything is physical, that is, that there is nothing over and above the physical. If this is so, then it is clear that physicalism is not committed to the claim that every token is identical to some physical token.[4] That is, physicalism is not committed to token-physicalism either.[5]

[4] See Stoljar (2005a) and Chalmers (1996: Ch. 2) for similar arguments and further discussion of type-physicalism and token-physicalism.

[5] However, the theses of type-physicalism and token-physicalism have played an important role in the discussion of physicalism in the second half of the last century. (See for instance Place (1956), Smart (1959), Putnam (1967), Davidson (1970) and Fodor (1974), among many others.) Some classical anti-physicalist arguments, such as Saul Kripke's Modal Argument or Stephen White's Property Dualism Argument, are really directed against type-physicalism and token-physicalism. Given the significance of these arguments, we will also discuss them in the following chapters, even if they are not, strictly speaking, arguments against *physicalism* (as we will characterise it here). In any case, those arguments can be seen as precursors of the arguments against physicalism (properly understood) that we will be mainly concerned with, and it will be interesting to see how they are related. We will say much more about this issue in Ch. 3.

Therefore, we can conclude that (I*) does not seem to provide a satisfactory answer to the completeness question. That is, the idea that everything in the actual world satisfies the condition of being physical is not really committed to the claim that all entities in the actual world are identical to physical entities (that is, those posited by physics). What is the core claim of physicalism then? What relation should hold between the entities posited by physics and all the rest of entities in the actual world, so that we could say that there is nothing over and above the physical?

A popular answer to this question is the following: everything is physical if and only if everything that exists somehow *depends on*, or is *determined by*, physical entities (i.e. those posited by physics). The idea here is that an entity or property can be determined by physical entities or properties, without being identical to them.

1.2.2. Introducing supervenience

The notion of *supervenience* aims to capture this intuitive notion of dependence, or determination. The crucial idea here is that there are some entities, facts, properties, etc. that depend on some other more fundamental entities, facts or properties (we will mainly focus on properties in what follows).[6] For instance, some people think that the aesthetic properties of a painting depend on the physical properties of the painting, that is, how the colours and textures are arranged, etc. The idea is that the physical properties of the painting determine the aesthetic properties of the painting, such as whether it is a beautiful painting, what it represents, what style it belongs to, and so on. We can use the notion of supervenience to express this idea: we will say that the aesthetic properties of the painting *supervene* on its physical properties.

What does that mean exactly? One way of cashing it out is the following: take the property of being a beautiful painting. Plausibly, if we have two paintings that are physically identical, then they will be equally beautiful. That is, two paintings that are identical in their physical qualities have to be identical with respect to how beautiful they are. In other words, if we have two paintings that differ in how beautiful they are, then they must differ in their physical properties too. If this is the case, we will say that the property of being beautiful *supervenes on* the physical properties of the painting.

[6] One reason to focus on properties is this: the claim that all properties in the actual world are physical arguably entails that all *particular entities* are physical as well, since these will only have physical properties. Therefore, if we define physicalism in terms of all properties being physical, physicalism will entail that all entities are physical, as intuitively required.

We can express this idea by means of the following schema:

(*Supervenience*): A-properties supervene on B-properties when two *entities* cannot differ in A-properties without differing also in B-properties.

That is to say, if two entities are different concerning A-properties, they have to be different concerning B-properties. And if they are identical concerning B-properties, they will also be identical concerning A-properties.

So, how can we answer the completeness question, using the notion of supervenience? A first approximation is this:

(II): Everything is physical if and only if all properties *supervene* on physical properties.

As we saw above, we can understand physicalism in terms of the claim that everything is determined by physical properties, that is, those posited by physics. We can express this idea in terms of supervenience: the thought is that the way all things are *supervenes* on the physical:

(*Physicalism II*): All properties supervene on physical properties.

Different characterizations of the notion of supervenience have been offered in order to provide a more precise characterization of physicalism.[7] In particular, one debate concerns what type of 'entities' are at issue in the (Supervenience) schema above. Two main notions that have been widely discussed are the notions of *local* and *global* supervenience:

(*Local Supervenience*): A-properties supervene on B-properties if and only if, for *any possible individuals* x and y, if x and y are identical concerning B-properties, they are also identical concerning A-properties.

[7] See Kim (1993) for a classical survey. See also McLaughlin (1995) and Stalnaker (1996).

(*Global Supervenience*): A-properties supervene on B-properties if and only if, for *any possible worlds v* and *w*, if *v* and *w* are identical concerning B-properties, they are also identical concerning A-properties.

The main idea of local supervenience is that there are some facts about an individual—the B-facts—that determine the A-facts about that individual, in the following sense: if two possible individuals are identical concerning B-properties, they will be identical concerning A-properties. [8]

The main idea of global supervenience is that there are some facts about the world—the B-facts—that determine the A-facts, in the following sense: if two possible worlds are identical concerning B-properties, they will be identical concerning A-properties.

These definitions involve the notion of *possible worlds*. Possible worlds are a theoretical tool that philosophers have introduced in order to talk about possibility and necessity. In this way, we can say that something is possible when there is a possible world where it is true, and that something is necessary when it is true in every possible world. A possible world is a complete way the world might be, where every detail is settled. For instance, the way the actual world is is obviously one way the world *might* be, so the actual world is a possible world. But there are other ways the world might be: for instance, France could have won the 2006 World Cup, instead of Italy. Then, there is a possible world that is identical to the actual one in all respects, with the exception of the fact that France is the winner of the 2006 World Cup in that

[8] This notion of supervenience is similar to what has been called 'weak supervenience', but should not be confused with it:

(Weak supervenience): A-properties supervene on B-properties if and only if, for any individuals *x* and *y in the actual world*, if *x* and *y* are identical concerning B-properties, they are also identical concerning A-properties.

This notion of supervenience is usually taken to be too weak for the purposes of defining physicalism. The reason is that the relation of weak supervenience might hold between properties that are just contingently related in the actual world, but where there is no interesting relation of dependence or determination. For instance, it might be a contingent fact about the actual world that all blond people enjoy listening to Madonna's songs. But it is clear that, even if that was true, the property of being blond does not determine in any interesting sense the property of enjoying listening to Madonna's songs, although it would be true in that case that the latter property would weakly supervene on the former, since all individuals in the actual world that were identical with respect to the property of being blond would also be identical with respect to the property of enjoying listening to Madonna's songs. For this reason, philosophers have rejected the notion of weak supervenience. The notion that I have called 'local supervenience' is also called sometimes 'strong supervenience' (for instance, in McLaughlin (1995)). I will use the label 'local supervenience', since the contrast with global supervenience is clearer in that way. See Kim (1993), McLaughlin (1995) and Stalnaker (1996) for more discussion of these notions of supervenience.

world. Here, by 'possible world' we mean a possible *universe*. There are some possible "worlds" in this sense (possible universes) that differ from the actual one with respect to facts outside the planet Earth. For instance, there is a possible universe where the only difference is that Venus is one metre closer to the sun than in the actual universe. We will call that possible universe a possible *world* too.[9]

1.2.3. A problem for local supervenience

We could define physicalism in terms of local supervenience, as follows:

> (III): Everything is physical if and only if all properties *locally supervene* on physical properties.
>
> (*Physicalism III*): All properties *locally supervene* on physical properties.

However, there is a problem with this characterization of physicalism. In particular, there is a problem with the idea that all properties of an individual supervene *locally* on the physical properties of that individual, since this does not seem to be the case for many properties. In any case, that is something that our characterization of physicalism should not be committed to. For instance, according to (Physicalism III), if any two subjects are physically identical, they are going to be mentally identical as well, independently of their environment. However, this characterization of supervenience is too strong, because arguably there are subjects that are physically identical, and in particular have identical brain configurations, but are located in different environments, and therefore have *different* mental properties. If this is correct, mental properties will not locally supervene on the physical properties of a subject.[10] One way of arguing for this is as follows. Arguably, the content of a mental state depends not only on the physical properties of the brain (or the whole individual) but also on the physical properties of a subject's environment. If this is so, then the

[9] The issue of the nature of these non-actual possible worlds is a very controversial one. I will try to remain as neutral as possible in what follows.

[10] Chalmers (1996: 33-34) provides a similar argument against the characterization of physicalism in terms of local supervenience.

mental states of two physically identical subjects placed in different environments might have different contents.[11]

One way of motivating this view is as follows. Plausibly, when you think about water in this world, your thoughts are about water, that is, H_2O. But now imagine that you are living in the hypothetical scenario devised by Putnam (1975), the so-called *Twin Earth*. In this planet, the odourless, colourless stuff that fills rivers and lakes and falls from the sky (the "watery stuff", for short) is not H_2O but XYZ. If you were living there, your thoughts would not be about water but rather about this other substance, which we can call 'twin-water'. Here we have a case of two possible individuals (you, and your counterpart in Twin Earth) who are physically identical, we can assume, but who nonetheless differ with respect to some of their mental properties: while your mental states are about H_2O, your counterpart's corresponding mental states are about XYZ.

If these considerations are correct, mental properties do not supervene locally on the physical properties of an individual. But the fact (if it is a fact) that mental properties depend on subjects' environments in this way should not be, *per se*, a reason to reject physicalism, and therefore physicalism should be characterised so as to be compatible with the failure of local supervenience of mental properties (and many other properties).

The notion of global supervenience is more appropriate because it is compatible with the context-dependent character of some properties. Even if, say, mental properties depend on their environments as we have suggested, they could still supervene globally on the physical, because it could very well be the case that any two physically identical worlds were mentally identical. In particular, if two worlds are physically alike, they will also be alike concerning the environments of subjects, and therefore, they can be alike concerning mental states.

1.2.4. A defence of global supervenience

We have seen that a characterization of physicalism based on local supervenience is too strong. There seem to be certain situations according to which mental properties do not supervene *locally* on the physical, but we would still say that in those situations

[11] This is the popular doctrine known as *content externalism*. For more discussion of this, see, for instance, Braddon-Mitchell and Jackson (1996: Ch. 12).

physical properties determine mental properties, if we take into account the physical properties of entities outside the individual's body.

Global supervenience is not too strong in that way, because it is compatible with the context-dependent character of mental properties (and other properties). However, some philosophers think that global supervenience is too weak to capture the intuitive idea behind physicalism and, therefore, too weak to characterize physicalism.[12] In order to examine their worries, let's introduce the corresponding answer to the completeness question:

> (IV): Everything is physical if and only if all properties *globally supervene* on physical properties.

The corresponding characterization of physicalism would be the following:

> (*Physicalism IV*): All properties *globally supervene* on physical properties.

The alleged problem for this new formulation of physicalism can be put in this way: There are certain scenarios in which physicalism, as intuitively conceived, seems to be false, but (Physicalism IV) still holds, that is, all properties globally supervene on the physical. In that case, we would have a counterexample to (IV): the proposition on the right-hand side of the biconditional would hold but the one on the left-hand side would not.

One of such possibilities is the following. Imagine a possible universe, W, that is almost physically identical to the actual one, molecule by molecule, the only difference concerns one atom on the surface of the recently discovered dwarf planet Eris. Say, in the actual universe there is a carbon atom at a certain location on the surface of Eris, whereas there is a silicon atom at that location in W. For all the thesis of global supervenience says, this alternative universe W could be dramatically different from the actual one concerning many properties, for instance mental properties: maybe there are no mental properties at all in W, or maybe they are completely different. Global supervenience just talks about possible worlds that are *physically identical*, molecule by molecule, to the actual word: these worlds have to

[12] See, for instance, Kim (1993: 277-8).

be identical in all respects, including mental properties. But for worlds that differ physically, even in very minor respects, the principle of global supervenience remains silent: any distribution of mental properties would be possible.

It could be argued that physicalism is incompatible with such a scenario and that therefore, a good definition of physicalism should rule out this possibility. If so, it would follow that global supervenience is too weak to capture the intuitive notion of physicalism. This could be offered as a reason to prefer the notion of local supervenience for that task. But I think it is not necessary to go that far. Robert Stalnaker (1996) and Daniel Stoljar (2005a) have suggested that, although it is the case that possibilities such as the one just described concerning the atom on the surface of Eris seem very implausible, physicalism does not have to rule them out on its own: we have additional evidence that allows us to exclude such far-fetched scenarios. The intuitive idea of physicalism is that *the physical* determines the mental (and all the other realms); physicalism does not say anything with respect to the further question of *which* physical properties in particular determine which mental properties. Physicalism is just committed to the claim that all properties supervene on physical properties. It should be neutral concerning which physical properties are responsible for which higher-level properties. For all physicalism says, a change in a single atom could have huge consequences. Physicalism does not have to exclude that: we have science, and common sense, to tell us what physical properties are responsible for the different instances of higher-level properties, and in particular mental properties, of the actual world.[13]

1.2.5. The problem of immaterial angels

A further and more worrying problem for the characterization of physicalism as (Physicalism IV) is the following. Imagine that there is a possible world W*, physically identical to the actual one (which we can denote @) molecule by molecule, where in addition there is some non-physical stuff, say, some immaterial angels. We can assume that these angels possess some kind of mental life, for instance, they might have mental states with intentional content. If there was such a possible world

[13] It could be argued, however, that the relation of global supervenience is too weak to capture any interesting notion of dependence or determination among, say, mental and physical properties. For instance, Melnyk (2003) argues that we have to supplement it with the notion of *realization*. In any case, as we will see in section 1.2.6, this is not a problem for our main purposes here, since our main aim is to show that physicalism, as intuitively conceived, is at least *committed* to a claim of global supervenience, even if it is not equivalent to that claim. For more on this, see 1.2.6 below.

W*, it seems clear that the global supervenience of mental properties on the physical would not hold. Recall the definition of global supervenience:

> (*Global Supervenience*): A-properties supervene on B-properties if and only if, for any possible worlds *v* and *w*, if *v* and *w* are identical concerning B-properties, they are also identical concerning A-properties.

In the case just described, there are two possible worlds that are identical concerning physical properties (B-properties), but not identical concerning mental properties (A-properties), since there are some mental states in W* which are not present at @ (namely, the angels' mental states). Therefore, mental properties would not globally supervene on the physical, and (Physicalism IV) would be false.

But this scenario does not concern the actual world: it is just a description of some *possible* world where there are non-physical entities such as angels, with non-physical mental states. Why should this mere possibility be a problem for the claim of physicalism, which concerns the *actual* world? For all our little story says, the actual world might very well be completely physical. In particular, what gives rise to the problem is the fact that @ and W* differ: W* contains, whereas @ lacks, non-physical entities. This seems to be compatible with the intuitive idea behind physicalism: the actual world lacks non-physical entities. How could this story, then, pose a problem for physicalism? Where did we go wrong?

The problem lies in our definition of physicalism: (Physicalism IV) is, after all, too strong, since it rules out the possibility that there are other possible worlds that contain non-physical stuff. That is, (Physicalism IV) rules out the possibility that there are other worlds where physicalism is false. But physicalism, at least intuitively, is compatible with there being other possible worlds where physicalism is false.

So how can we avoid this problem? The first step is to make clear that physicalism is a claim about our actual world. We can still use the notion of global supervenience, as follows:

> (V): Everything is physical if and only if, for any possible world *w*, if @ (the actual world) and *w* are identical concerning physical properties, they are also identical concerning *all* properties.

(*Physicalism V*): For any possible world *w*, if @ (the actual world) and *w* are identical concerning physical properties, they are also identical concerning *all* properties.

In other words, this new characterization of physicalism basically says that physicalism is true just in case any physical duplicate of the actual world is a duplicate *simpliciter*, that is, a duplicate concerning all properties.

However, this new definition has not completely solved the problem. In fact, it still suffers from very much the same problem: a possibility such as the one described above (W*), concerning non-physical angels, also constitutes a counterexample for the claim (Physicalism V). For the possible world W* is a physical duplicate of the actual world, @, but it is not a duplicate *simpliciter*, that is, they are not identical concerning all properties. At least, they differ with respect to the property 'containing angels'. And plausibly, they will differ in the number of token mental states they contain: W* will contain some additional mental states, due to the presence of some very thoughtful angels.

Again, the problem here is that our definition of physicalism rules out possibilities that physicalism should not rule out, namely, the existence of possible worlds that are physically identical to the actual world and that in addition contain some non-physical entities, which could instantiate some new mental properties. Physicalism should be compatible with this being a possibility, although not with this being actual. One solution, suggested by Frank Jackson (1994), is to reformulate our definition of physicalism along the following lines:

(VI): Everything is physical if and only if any possible world *w* which is a *minimal* physical duplicate of @ is a duplicate *simpliciter*, i.e. a duplicate of @ concerning all properties.

(*Physicalism VI*): Any possible world *w* which is a *minimal* physical duplicate of @ is a duplicate *simpliciter*, i.e. a duplicate of @ concerning all properties.

What is a minimal physical duplicate of the actual world? The answer is simple: a duplicate of the actual world that contains all the physical entities, properties and relations that exist at the actual world, and *nothing else*. Imagine that you had the

power to create new possible worlds: how would you go on to create minimal physical duplicates? You would have to add all the physical entities of the actual world, molecule by molecule. (Remember that physical entities are those posited by physics.) You could not miss a single physical entity. And, once you add all the physical entities, you *stop there*: you cannot add anything else. We can think of it in this way: our recipe for making minimal duplicates would consist of a complete physical description of the actual world, so that we have to add everything that is listed there; and at the end of that very long list, we would have a special warning: a 'that's all' clause (preferably in red letters), so that we are not allowed to add anything extra. In this way, you would get a minimal physical duplicate of the actual world.

So, does this new characterization of physicalism solve the problem of the previous one? Do minimal physical duplicates help at all? Well, let's see what happens with the possible world described above, namely W*. Does (Physicalism VI) rule it out? It seems it does not, because W* is clearly not a minimal physical duplicate of the actual world, since it contains a lot of extra entities beyond the physical ones. So it does not satisfy the constraints for being a minimal physical duplicate. Thus, (Physicalism VI) does not say anything about W*: it is completely neutral about what kind of entities populate W*. (Physicalism VI) is purely concerned with minimal physical duplicates of @: these have to be duplicates *simpliciter*, so they should not contain additional non-physical angels, or else they will not be identical to the actual world concerning all properties. But (Physicalism VI) is compatible with W* containing as many mental properties as we want, since it is not a minimal physical duplicate.

So (Physicalism VI) gives the right result: it does not rule out the possibilities that physicalism should not, intuitively, rule out. (Physicalism VI) looks like the best characterization of physicalism that we have got so far.[14] This characterization of physicalism is widely held in contemporary philosophy of mind.[15] More importantly, that is the definition of physicalism that the advocates of the anti-physicalist arguments that we will examine in this thesis are concerned with. That is the claim that they aim to attack. From now on, we will call it simply 'Physicalism':

[14] Although it also has some detractors. We will see some of the alleged problems in the following sections of this chapter.
[15] Some prominent advocates of this characterization of physicalism are Lewis (1994), Jackson (1994) and (1998a), and Chalmers (1996).

> **Physicalism**: Any possible world *w* which is a *minimal* physical duplicate of @ is a duplicate *simpliciter***.**

The proponents of this characterization, then, endorse the following biconditional:

> (VI): Physicalism (as intuitively conceived) is true if and only if **Physicalism** is true.

In the following section, we will explore some further issues concerning this biconditional, and we will conclude that we have good reasons to endorse the claim that if physicalism as intuitively conceived is true, then Physicalism is true (although as we will see, the converse implication is more controversial).

1.2.6. Physicalism: necessary and sufficient conditions

To further motivate our characterization of physicalism, we can examine whether it provides necessary and sufficient conditions for the truth of physicalism. That is, we will first examine a situation in which Physicalism fails, to see whether physicalism would intuitively fail as well; and secondly we will examine a situation in which physicalism as intuitively conceived fails, to see whether Physicalism would fail too. We will see that the former question is much more straightforward than the latter.

So let us assess the issue of the *necessity* of the characterization first. That is, let us examine the biconditional (VI) in the left-to-right direction: 'If physicalism is true, any possible world *w* which a minimal physical duplicate of @ is a duplicate *simpliciter*'. We will consider a situation in which the consequent fails. For instance, we can imagine a possible world *V* which is a minimal physical duplicate of the world but not a duplicate simpliciter, because, say, *V* differs from @ with respect to some mental properties. How could this be? Could *V* contain some additional non-physical entities, which instantiate some mental properties not instantiated in the actual world? Well, this cannot be the case since by hypothesis *V* is a minimal physical duplicate, and this means that it does not contain anything besides the physical entities, properties and relations that are instantiated in the actual world. *V* is physically identical to @ and it contains nothing else, so if *V* and @ differ, @ must contain some entities or properties that do not appear in *V*. But *V* is a minimal physical duplicate, so by definition it contains all the physical entities that appear in @. So if there is

something that @ has and *V* lacks, it cannot be physical. Thus @ contains something that is non-physical, be it an entity, a property or a relation, and physicalism is false at the actual world.

Therefore, if the right-hand side of (VI) fails, then physicalism intuitively fails too. That is, Physicalism successfully provides *necessary* conditions for the truth of physicalism: if those conditions fail, physicalism will fail too.

What about the *sufficiency* of such conditions? This is a more difficult question. What we are examining now is the right-to-left direction in the biconditional (VI), that is: 'If any possible world *w* which is a *minimal* physical duplicate of @ is a duplicate *simpliciter*, then physicalism is true'. Does this fit our intuitive notion of physicalism? Is there any scenario where physicalism intuitively fails but it is still the case that any minimal physical duplicate of @ is a duplicate in all respects? It may seem, at least at first sight, that that is impossible. Let us assume that physicalism is false: then there are some entities or properties in the actual world which are not physical. Then, we would have minimal physical duplicates of the actual world that lacked such non-physical entities, since the only entities that we are allowed to put in a minimal physical duplicate are physical entities. So, if physicalism is false, there will be some minimal physical duplicates of @ that are not duplicates of @ in all respects, since @ will contain some non-physical entities that the minimal physical duplicate will lack.

If this is correct, it seems that our characterization of physicalism does provide sufficient conditions for physicalism to be true. Does it? One problem that has been suggested[16] is the following: there seem to be some clearly non-physicalist views about the mind which might be compatible with Physicalism. According to one such view, the mind is a non-physical entity, like a non-physical Cartesian soul, which is nonetheless necessarily connected to the physical body, so that every physical duplicate of the body will necessarily be accompanied by its corresponding non-physical soul. If this was the case, then it seems that mental properties would globally supervene on the physical, since any physical duplicate of the world would contain the same souls, and therefore, the same mental properties. In particular, minimal physical duplicates of the world would also contain such non-physical souls. How is this possible? Well, what our recipe for making minimal physical duplicates says is

[16] See, among others, Jackson (1998a: 22-23) and Stoljar (2005a).

this: 'add any physical entity that exists in the actual world. And stop there'. This clause would stop us from adding any additional stuff. But what if there were some entities that were automatically added by virtue of putting certain physical stuff in such a world? According to the "soul" view just described, the relation between our minds and our bodies is exactly like that: the connection between mind and body is impossible to break. This means that there is no possible world which is physically identical to but mentally different from the actual world, not even a minimal physical duplicate, since this would have to contain the same minds as the actual world.

Therefore, there seems to be one view according to which every minimal physical duplicate of the actual world is a minimal duplicate *simpliciter*, but nonetheless physicalism, as intuitively conceived, is clearly false, since according to such a view there exist lots of non-physical souls. If such a view is a coherent one, then it seems that Physicalism does not provide sufficient conditions for physicalism to be true, since Physicalism can be satisfied without physicalism being true.

In order to generate this problem, we read the definition of 'minimal physical duplicates' in a certain way: minimal physical duplicates are the result of adding every physical entity of the actual world and stopping there. This allows the possibility that there are minimal physical duplicates of @ which contain some additional entities not added by the "creators" of such possible worlds but automatically added, so to speak, given the nature of the connections between the physical and the non-physical entities. Could we solve this problem by imposing a stricter reading of the recipe for making minimal physical duplicates? In particular, let us examine what happens when we read the recipe as follows: a minimal physical duplicate is a possible world which contains every physical entity, property and relation instantiated in @, and *nothing else*. This new definition does not say what we can and cannot add to a minimal physical duplicate but rather what these minimal duplicates actually contain.

So let's revisit the problem with this new notion of 'minimal physical duplicate' in mind. If the dualist view described above (the one according to which minds are non-physical souls necessarily connected to our physical bodies) was true, what would follow? Well, if such view was correct, then there would be no minimal physical duplicates of the actual world. They would be plainly impossible, since every physical duplicate of the world would contain some non-physical stuff, due the necessary connections between physical bodies and souls. So there would be no

minimal physical duplicate of the actual world. Thus, it would still be the case that any minimal physical duplicate of the actual world is a duplicate simpliciter. This would be trivially true, since there are no minimal physical duplicates at all.

It seems, then, that we are stuck with this problem. What seems to cause it is that our characterization of physicalism seems to presuppose that there are no necessary connections between physical and non-physical entities. That is, our characterization of physicalism seems to presuppose that if an entity is non-physical, then it will be possible to have a possible world which is physically identical to @ but lacks the corresponding non-physical entity. This is the idea underlying our account of physicalism.

This seems an application of what is known as 'Hume's Dictum', according to which "there are no necessary connections between distinct existences" (Stoljar (2005a: 5)). It seems that our characterization of physicalism presupposes something along these lines. So we can assert that if Hume's Dictum is correct, and physicalism as intuitively conceived is *false*, then there will be *some* minimal physical duplicate of the world that is not a duplicate *simpliciter*. For, if Hume's Dictum is correct, then the dualist view described above is just incoherent: non-physical minds could not be necessarily connected to physical bodies in such a way. If there were non-physical souls in the actual world, there would be some minimal physical duplicate of @ that would lack them.

So it seems that if we want (VI) to provide sufficient conditions for physicalism to be true, we have to endorse Hume's Dictum. At any rate, what our characterization of physicalism is really presupposing is that what makes an entity non-physical (that is, the kind of entity that would make physicalism false) is that it is not necessarily connected to the physical. If this is rejected, then our characterization of physicalism will not work.

In any case, to finish this discussion, it is important to bear in mind that this issue poses a problem only for the *sufficiency* of our account. That is, the problem is whether it is sufficient, for physicalism to be true, that every minimal physical duplicate of @ is a duplicate simpliciter. It is clear, as we saw earlier, that this condition is *necessary* for physicalism to be true. That is, the thesis of physicalism is committed to there being no minimal physical duplicate of @ which is not a duplicate simpliciter. As we saw before, if there is a minimal physical duplicate of @ which is not a duplicate simpliciter, it is obvious that physicalism will be false. So if the

opponents of physicalism could establish that there are such minimal physical duplicates which are not duplicates in all respects, they would have established that physicalism is false.

This is precisely the strategy that the anti-physicalist arguments that we will be concerned with will follow. So we can see now what motivates this strategy. We have captured some widely agreed-upon necessary conditions for physicalism to be true. Maybe something else must be added to that account in order to turn it into a sufficient account.[17] In any case, if those necessary conditions are shown not to hold in the actual world, physicalism will be in trouble. And assessing whether or not they hold will be the main bulk of this work.

1.3. The Condition Question: Defining the 'Physical'

In the discussion so far we have focused on the *completeness* question, which considers what it is for *everything* to be physical. The best answer we have come up with is the claim we have labelled 'Physicalism': everything is physical if and only if every minimal physical duplicate of the actual world is a duplicate simpliciter. In this section, we are going to discuss the other important issue involved in the question of what it takes for physicalism to be true in the actual world. This second question is the so-called *condition* question, and it asks what it is for something (or for everything, in this case) to be *physical*. As I mentioned before, we have been assuming so far a provisional characterization of 'physical', according to which an entity, property or relation is physical just in case is has been posited by physics.[18] With this understanding of 'physical' in mind, we have proceeded to discuss the completeness question.

This notion of the physical provides some intuitive understanding of what it is for something to be physical, but on reflection, there are some problems that might affect our characterization of physicalism. We will explore some of these problems in this section, and we will try to provide a solution, so as to defend our characterization of physicalism.

---

[17] For instance, as we saw in section 1.2.4, it could be argued that global supervenience claims are too weak to capture the intuitive idea of physicalism. However, this is compatible with the claim that physicalism is committed to a global supervenience claim (namely, Physicalism), which is all that is required by the anti-physicalist arguments that we will explore.

[18] As before, we will mainly focus on properties in what follows.

Before doing that, some questions of clarification are in order. We have said that a property is physical just in case it is posited by physics. On the other hand, we have said that everything is physical just in case everything globally supervenes on the physical (in the sense specified in (Physicalism) above). Then, it seems that we can obtain two different notions of 'physical': on the one had, we have a notion of 'physical' according to which physical properties are those posited by physics. We will call this the *primary* notion of the physical (Dowell (2006)), or physical in a *narrow* sense (physical_n) (Melnyk (2003)). On the other hand, we have a broader notion of 'physical', according to which a property is physical just in case it bears a certain relation to physical_n properties. This is what is known as the *secondary* notion of 'physical' (Dowell (2006)), or physical in a broad sense (physical_b) (Melnyk (2003)). In this way, physicalism amounts to the claim that all properties are physical in the *broad* sense.

With this new terminology, the problem of defining physicalism can be easily explained: on the one hand, we have to provide a characterization of 'physical_n', that is, physical in the primary or narrow sense. This is not enough, though, because it is pretty clear that not all entities and properties are physical_n. So, the second task is to extend the class of properties that are acceptable for the physicalist, that is, the kind of properties that can exist in the actual world without making physicalism false. This is, in essence, the completeness question that we have examined earlier. Our characterization of 'physical' in the secondary, broad sense relies upon the notion of supervenience: a property *F* is physical in the *broad* sense (physical_b) if and only if any minimal $\textit{physical}_n$ duplicate of the actual world is a duplicate with respect to *F*. Thus, our official characterization of physicalism involves the narrow sense of 'physical', as follows:

> **Physicalism**: Any possible world *w* which is a minimal physical_n duplicate of @ is a duplicate simpliciter.

Our task now is to examine the narrow notion of 'physical' that we have been using so far, in order to see whether it gives rise to any significant problem that could jeopardize our characterization of physicalism.

(i) *F* is physical_n if and only if *F* is posited by physics.

One problem with this definition is the following: if we understand it literally, it could include any property that has been posited by any physical theory that has been proposed in the history of physics. But we do not want properties such as *phlogiston* count as physical$_n$. So it will be useful to restrict our definition to the properties and entities posited by the current physical theory, that is, those entities that are nowadays included in an account of the fundamental nature of our world. So the new proposal is this:

(ii) *F* is physical$_n$ if and only if *F* is posited by current physics.

1.3.1. Hempel's dilemma

This new definition faces another more serious problem. This problem is what is known as 'Hempel's dilemma' (see e.g. Hempel (1969)). The problem can be put as follows: when we try to define the narrow notion of 'physical', we face a dilemma. Either we appeal to current physics, as we have done here, or we appeal to future, ideal physics. In either case, the argument goes, there are serious problems that threaten the adequacy of our characterization of physicalism. Let's see why.

The first horn of the dilemma goes like this. If we define physical$_n$ as we have done in (ii), then – the argument goes – the corresponding characterization of physicalism would be false, even in situations that should not falsify physicalism, intuitively. The reason is this: through the history of physics, there are many physical theories that have been proven wrong, and many new theories have been proposed. Some entities and properties that used to figure in the accepted theories were later rejected, and some new entities have been introduced. By induction, there are good reasons to think that our current physics is not complete: plausibly, new entities and/or properties will be introduced by future physics. Let's consider one of these new entities, call it X. Now, if we define physical$_n$ in terms of current physics, then a minimal physical$_n$ duplicate will be any possible world that is identical with respect to all the entities and properties posited by current physics. That means that a possible world that is identical in everything with the exception that it lacks X will be a minimal physical duplicate of the actual world. However, it is clear that this minimal physical duplicate is not a duplicate simpliciter, since it lacks X, whereas the actual world does contain X (by hypothesis). So, in this case, Physicalism would be false.

But it does not seem that physicalism, as intuitively conceived, should be false in this scenario: the mere fact that physics has found a new kind of entity does not mean that there is something over and above the physical. For this new entity could very well be physical, in the intuitive sense. So if we define physical$_n$ in terms of current physics, the corresponding definition of physicalism will be inadequate.

So, what about defining physical$_n$ by appealing to future physics? This new definition might go like this:

(iii) *F* is physical$_n$ if and only if *F* is posited by future, ideal physics.

The idea here is to characterize physical entities or properties as those that are posited in an ideal physical theory that could be achieved in the future. This physical theory, by definition, would provide a complete and exhaustive account of the basic nature of the world: it would tell us about all the basic ingredients of the universe, without leaving anything out. Unfortunately, this new proposal also faces a serious problem (and this is the second horn of Hempel's dilemma). The problem here is that, if we define 'physical$_n$' as (iii) suggests, then the corresponding definition of physicalism would be trivially true. There would be no way the actual world might be that would make Physicalism, as characterised above, false. In particular, Physicalism would still be true in scenarios where it is obvious that physicalism, as intuitively conceived, would be false. For instance, let's imagine that the basic ingredients of paradigmatic physical objects, such as rocks or puddles, include some fundamental psychological properties. That is, let's imagine that the best account of the behaviour of rocks and puddles posits some psychological properties which are basic and fundamental, and do not depend on any other more basic properties of such physical entities. If this was the case, it seems obvious that physicalism, as intuitively conceived, would be false. But Physicalism, understood in terms of the physical$_n$ as defined in (iii), would still be true. For the ideal physical theory would include those fundamental psychological properties as fundamental properties of reality, since, by definition, this ideal physical theory is a complete theory and therefore it cannot leave anything out. So our corresponding notion of physicalism would be trivially true: any minimal physical duplicate of @ would contain any new entities that were introduced by the ideal physical theory. So any minimal physical duplicate of @ would be a duplicate simpliciter, even if the actual world contained entities or properties that were

completely different from the physical entities and properties that we recognise today. That is, even if the actual world contained entities or properties that are clearly non-physical. So, on this horn of the dilemma, we would also arrive at a clearly incorrect definition of physicalism.

The problem raised by Hempel's dilemma is quite difficult to solve, and it has generated a lot of discussion.[19] We do not have space here to discuss all the possible solutions, but I would like to point out one way of solving the dilemma that seems plausible to me. The idea behind the dilemma is that there is an intuitive sense of 'physical' such that, on the one hand, there might be unknown entities that are physical in that sense, and on the other hand, the actual world might turn out to contain basic ingredients that are not physical in that sense. The dilemma seems to presuppose this notion of 'physical', and this is what causes the problem, since neither (ii) nor (iii) as stated above can capture this notion of the physical. I think that the best solution is to accept these intuitions at face value, and to assume that there is some intuitive notion of 'physical' that is at work in these examples. It might be difficult to define, but it does not mean that the notion itself is incoherent.[20]

Furthermore, I think that some steps can be made towards a definition of such an idea, even if we are far from having a completely satisfactory definition. As Jackson and Braddon-Mitchell point out, "the hope is that the problems in current physical sciences are not going to call for the acknowledgement of properties, entities and relations different in kind from those now on the scene. The incompleteness of our current physical theory does not imply incompleteness in the *kinds* of ingredients that will be needed to complete the job" (1996: 13-14). That is, the hope of the physicalist is that future physical theory is not going to introduce entities and properties of a completely new kind. This is, after all, what physicalism claims: there is nothing over and above the entities and properties of the sort that current physics tells us about. With this intuitive understanding of 'physical', we allow that future physics may posit new entities and properties, but we put some limits upon the kind of entities that future physics might introduce, compatibly with physicalism. Admittedly, this understanding of 'physical' is vague and imprecise, but this corresponds with our

[19] For some recent discussion see, for instance, the discussion in Dowell (2006), and also, in the same volume, Wilson (2006) and Worley (2006).
[20] Stoljar (2005a: 13) offers a similar argument.

intuitive notion of the physical, which is equally vague and imprecise. So, maybe this characterization is not so bad, after all.

So we have arrived at this definition:

(iv) *F* is physical$_n$ if and only if *F* is a property of the same kind as the properties posited by current physics.

Accordingly, our characterization of physicalism, namely, Physicalism, should be understood in terms of this notion of 'physical$_n$'. In this way, physicalism will be true in the actual world only if any minimal physical duplicate of it (that is, any possible world identical to @ with respect to all the entities and properties of the same kind as those posited by current physics) is a duplicate in *all* respects.

This characterization of physicalism captures the minimal commitments of physicalism. It is not completely clear, as we have seen earlier, whether this characterization captures all the tenets of a physicalist view of the world. Nonetheless, what seems pretty clear is that if Physicalism, as characterized above, turned out to be false, then the actual world would be quite different from how the physicalist takes it to be. Therefore, any initially persuasive argument that concluded that physicalism in our sense is false should be taken seriously.

1.4. The Causal Argument for Physicalism

As we have just seen, we have arrived at a characterization of physicalism that captures the core claim of the physicalist worldview. As I have emphasised, any compelling argument against physicalism in this sense would put the physicalist view in jeopardy, since those arguments would be attacking a central claim of physicalism, a claim which virtually all physicalists are committed to. In this thesis, we are going to critically examine some of the most important arguments in that direction in great detail. Before turning to that task, however, it would be useful to say something about what makes physicalism plausible. We can agree that arguments against the claim that any minimal physical duplicate of the actual world is a duplicate simpliciter, if successful, would be very damaging to physicalism. But why would this be such a terrible consequence? Why should we believe in physicalism in the first place? What would be so bad about rejecting physicalism?

There are different ways of motivating physicalism, which have been discussed in great detail in the recent literature on philosophy of mind, so I will not go into all of them here.[21] But I would like to mention what I take to be one of the most powerful arguments for physicalism that is available. This argument is the so-called *causal argument*, and it gives us a very good idea of why rejecting physicalism would put some of our most deeply-entrenched beliefs in jeopardy.

The main idea behind the causal argument is, roughly, that we take many properties that are not physical_n (i.e. not the kind of properties posited by current physics) to be causally efficacious with respect to the physical_n realm. And in order to be causally efficacious, as I will explain later, those properties have to be at least globally supervenient on physical_n properties (that is, they have to be physical in at least the broad sense, i.e. physical_b). That is, it is widely accepted that a property can be causally efficacious with respect to physical_n properties only if it is physical_b, that is, only if it supervenes globally on the physical_n.[22] (I discuss this question a bit further below.)

As we have seen, physicalism is a very general thesis: it is concerned with all the entities and properties in the actual universe, and it says, of all of them, that they are physical: they are determined by the physical_n realm. We have seen that the best way to capture this idea more precisely is by means of the idea of global supervenience: every minimal physical_n duplicate of @ is a duplicate with respect to *all properties*. So a successful argument for physicalism would have as its conclusion the claim that all properties globally supervene on the physical_n, in that sense. The causal argument could be construed in this way, as a general argument about what it would take for any property to causally affect the physical_n.[23]

In any case, since the arguments that we are going to examine in the following chapters focus on the case of the conscious properties of mental states, I will first present the causal argument concerning those conscious properties. That is, I will focus on a version of the causal argument which argues that all conscious properties are physical_b. We will see that there are good reasons for holding that conscious properties are physical in that sense, and therefore, there are good reasons for

[21] For instance, Stoljar (2005a: 20-22), Melnyk (2003: 76-8) and Levine (2001: 21-38) discuss several arguments for physicalism. All of them consider the so-called 'causal argument', which I will explain in what follows, to be one of the main arguments for physicalism.
[22] See the compelling discussion of this issue in Papineau (2002: 32-36).
[23] At the end of this section, I present a general argument of this form.

examining very carefully any arguments that allegedly show that conscious properties are not physical$_b$.

This version of the causal argument can be put as follows:

(1) Conscious mental events have physical$_n$ effects, by virtue of their conscious properties.
(2) All physical$_n$ effects are fully caused by physical$_n$ events, in virtue of their purely physical$_n$ properties.
(3) The physical$_n$ effects of conscious mental events are not always over-determined by distinct causes.
(4) Conclusion: conscious properties are physical$_b$ properties.[24]

Let's examine each premise in turn. The first premise says that our conscious mental occurrences are causally efficacious, and in particular, that they cause physical events to be instantiated. This is a very familiar phenomenon: for instance, my conscious mental event of being in pain causes me to move to the kitchen and take an aspirin, which is a physical event. That mental state causes the physical event by virtue of instantiating the conscious property *being in pain*: after all, I went to take an aspirin because of the *pain* I felt. Another example is this: my visual experience of seeing the red traffic-light causes me to stop my car. Again, a conscious mental occurrence causes a physical event by virtue of the former instantiating a conscious property, i.e. the property of being the experience of seeing a red light. It is perfectly usual to invoke such causal connections: we typically believe that our conscious properties have plenty of physical effects. Our ability to communicate our conscious mental states presupposes this causal link: we are able to communicate our conscious mental events because they *cause* our verbal reports of them.

Therefore, the first premise is widely supported by common sense. It would be a great cost to reject the causal efficacy of our conscious properties in the physical world.

The second premise says that for any physical effect, there is a physical cause which can be identified in purely physical terms. This is the so-called 'completeness of physics': for any physical event, we can find a sufficient physical cause; we do not

[24] This presentation of the argument is based on the discussion in Papineau (2002: Ch. 1).

have to go outside the physical realm to explain the causal history of such a physical event. This claim is strongly supported by the findings of contemporary physics. (For a detailed defence, see Papineau (2002: appendix).) Therefore, rejecting this premise would be a great cost too.

Finally, the third premise says that we cannot explain every case of a conscious mental state causing a physical event by invoking an over-determination of that physical event. That is, we cannot explain the causation of such physical event in terms of it being caused by two distinct events at the same time (or two different properties of the same event). For instance, consider my mental state of being in pain which caused me to take an aspirin. This physical event is caused by the conscious property of being in pain. But, according to premise (2), this physical event has a purely physical cause, that is, we can find a physical property that is causally responsible for it. Plausibly, this physical property is some neurobiological property of my brain. So we have a physical effect with two distinct causes: the pain and the neurological state of my brain. Should we say that the physical effect of taking an aspirin was causally over-determined? A paradigmatic case of causal over-determination is, for instance, a person who is shot by two arrows, at exactly the same time, such that either of them would have caused the death. We can say in this case that this person's unfortunate death was over-determined by two distinct causes. But this kind of situation is very rare: it is not often the case that a physical event is over-determined like that. In particular, it would be very awkward to say that every time a conscious mental event causes a physical event, there is a case of over-determination involved. This would make over-determination a very common phenomenon, which it clearly is not. This is what motivates the third premise of the causal argument.

Therefore, what is the proper conclusion of the causal argument? The conclusion is that the conscious properties that are responsible for, say, my taking an aspirin, are not *distinct causes* from the physical properties that are responsible for my taking an aspirin. In this way, we can say that both properties were causally responsible, but the physical effect is not over-determined, since they are not really two separate causes. This leads us to the conclusion that the conscious properties are physical after all.

There are different ways of spelling out this idea that conscious properties are not distinct causes from the corresponding physical properties. A first option is the idea that all conscious properties are *identical* to some physical_n property or another,

but this seems too strong. Arguably, most mental properties are not identical to any physical_n property, since they can be realized by more than one physical_n property. Hence, if we hold that a conscious property can causally affect the physical_n only if it is identical to some physical_n property, this would deprive most conscious properties of causal efficacy with respect to the physical_n realm. Therefore the proper conclusion of the causal argument has to be that conscious properties are physical in a *broader* sense, and a very plausible option here is to appeal to the notion of 'physical_b'. (Recall that a property *F* will be physical_b in this sense when it globally supervenes on the physical_n, or more precisely, when any minimal physical_n duplicate of @ is a duplicate concerning *F*.)

However, some philosophers have argued that being physical_b does not guarantee the causal efficacy of a property. They argue that globally supervening on the physical is not sufficient for a property *F* to be causally responsible for a given physical_n effect *G*, since the physical_n properties that are also causally responsible for *G* can pre-empt the physical_b property *F* from being causally efficacious.[25] The discussion of this issue in contemporary philosophy of mind is very rich and sophisticated, so we do not have the space to examine it in any detail here. In any case, I would like to note that this issue does not really matter for our purposes. The main consequence of the causal argument, for the purposes of our discussion, is that conscious properties have to be *at least* physical_b in order to causally affect the physical. It is not clear whether being physical_b is *sufficient* for warranting their causal efficacy, but it is widely held that being physical_b is a *necessary* condition for being causally efficacious with respect to the physical_n. If a property does not globally supervene on the physical, in the appropriate sense, then it is dubious that it can exert any causal powers with respect to the physical_n realm. Therefore, the proper conclusion of the causal argument for physicalism concerning conscious properties is that conscious properties, since they have physical_n effects, must be at least physical_b properties. And this is a pretty strong consequence.

We can also notice that, if the causal argument generalises, that is, if it can be applied to any kind of existing property, then we can obtain a general argument for Physicalism, as characterised above. The argument can be put as follows:

---

[25] See, for instance, the excellent discussion in Kim (1998). Kim argues that the notion of supervenience, by itself, does not solve the problem of the causal efficacy of mental properties but rather gives rise to it.

(1') Non-physical$_n$ properties are causally responsible for the instantiation of some physical$_n$ properties.[26]

(2') For any physical$_n$ effect, there is a purely physical$_n$ cause.

(3') Physical$_n$ effects are not always over-determined by two distinct causes.

(4') Conclusion: Non-physical$_n$ properties must be identical to physical$_b$ properties.

The evidence in favour of these new premises is very similar to the evidence for the previous formulation of the argument. In this case, the conclusion claims that all real properties that are non-physical$_n$ must be (at least) physical$_b$. That is, the conclusion asserts that all properties in the actual world are physical$_b$. And, according to the definition of 'physical$_b$' above, this amounts to saying that every minimal physical$_n$ duplicate of the actual world is a duplicate concerning all properties. This corresponds perfectly with our characterization of Physicalism above. So the proper conclusion of the causal argument, in the generalised version, is that Physicalism is true.

[26] This is a very strong claim: unmodified, it says that *all* the properties that we accept in our ontology have to be causally responsible for the instantiation of some physical$_n$ properties. This can seem controversial, for instance with respect to abstract properties or evaluative properties. With respect to the case of *abstract* properties, one solution would be to claim that physicalism is restricted to a claim about the nature of *concrete* entities and properties. With respect to *evaluative* properties, it is not clear to me what to say, but perhaps it could be argued that if evaluative properties are not causally efficacious with respect to the physical$_n$, then they are not genuine properties.

2. Epistemic Arguments against Physicalism I: Conceivability Arguments

2.1. Phenomenal Consciousness

Consciousness is a very familiar phenomenon. For most of our waking lives, we are enjoying conscious mental states. Most of these conscious mental states have a very striking feature: there is something *it is like* to undergo them;[27] they have some particular qualitative feature or subjective *feel*. In order to illustrate this aspect of consciousness, it will be useful to consider some examples. We can focus on familiar sensations, such as a toothache, or the pleasant experience of tasting Belgian chocolate, or the visual experience of seeing a sunset. All these experiences have something in common: they are *like something to us*. And we can distinguish them by what they are like.

This is the notion of consciousness that we will be concerned with here, and it is typically called *phenomenal consciousness*. We can say that a subject is phenomenally conscious in this sense when it is like something to be that subject. And we can also say that a mental state is phenomenally conscious when it is like something to be in that mental state. Phenomenally conscious mental states (or for short, *phenomenal states*) are characterised by what they are like, that is, by their qualitative features, also called *phenomenal properties* or *qualia*.

I hope I have said nothing very controversial so far. That there is such an aspect of consciousness is a shared view among the majority of physicalists and anti-physicalists alike.[28] In any case, it seems hard to deny that our conscious mental states are like something for us, that they feel in a particular way. The proper characterization of these phenomenal states and their phenomenal properties is a more complicated issue. One important question is about the *metaphysics* of such conscious mental states: Are they physical or not? Are the phenomenal properties instantiated by them physical or non-physical? There are also important and vexed *epistemological* issues concerning phenomenal consciousness. Some of these questions concern the

[27] This phrase comes from the classic Nagel (1974).
[28] However, of course, this being a philosophical debate, not everyone agrees. Daniel Dennett (1988), for instance, has famously argued that the notion of 'qualia' is incoherent and that nothing actually satisfies it.

possibility of offering an explanation of phenomenal consciousness, such as: Can we explain phenomenal consciousness in physical terms? Could we provide a reductive explanation of phenomenal consciousness? What elements are needed in a theory in order to explain phenomenal consciousness satisfactorily? Some other epistemological questions concern the way we come to know about our own phenomenal states: Can we know what phenomenal states we are having just by introspection? Could we come to know what phenomenal properties are instantiated by subjects other than ourselves? Finally, but not less importantly, there are also *semantic* questions in this area. We use bits of language to refer to our phenomenal states and properties: expressions such as 'pain' or 'headache' or 'seeing something red' are just some examples. These expressions raise some interesting and difficult questions, such as: What is the meaning of these terms? How do these terms get to refer to the phenomenal states and kinds they refer to?

My main focus in this thesis will be on the metaphysical questions involving phenomenal consciousness. In particular, I will be exploring the question of whether phenomenal consciousness is physical or not—in other words, the question of whether phenomenal states and properties are physical or not. In the previous chapter I have explained what this question means in some detail. As we have seen, the ontological claim of physicalism asserts, intuitively, that everything is physical, or more precisely:

> **Physicalism** is true if and only if any possible world *w* which is a *minimal* physical duplicate of @ is a duplicate *simpliciter*.

The main question that will concern us can be put easily: Does phenomenal consciousness pose a problem for physicalism? Does the existence of phenomenal states and properties make the thesis of physicalism, so characterised, false? This is tantamount to asking whether phenomenal states and properties are physical or not. Of course, physicalism could be false for many reasons: any existing non-physical property would make it false. Here, we aim to know whether phenomenal consciousness in particular makes it false. If we conclude that it does not, we have not established yet that physicalism is true, of course, but we would have taken a very important step in that direction, since phenomenal consciousness poses one of main contemporary challenges to physicalism.

As we have seen in the previous chapter, when we ask whether phenomenal properties are physical or not, we are not asking whether phenomenal properties are the kind of properties that are posited by current physical theory (that is, whether they are $physical_n$ properties), because if that were the question, the answer would be obviously 'no'. (Unfortunately, our question does not have such an easy answer, so that means there is more work for us to do.) Rather, the question that we are interested in is whether phenomenal properties (and states) are physical in a broader sense, which we defined in the previous chapter as properties that globally supervene on the physical, or more precisely:

> A property *F* is physical in the *broad* sense ($physical_b$) if and only if any minimal $physical_n$ duplicate of the actual world is a duplicate with respect to *F*.

So our task will be to examine whether phenomenal properties are physical in this sense, or more generally, whether the existence of phenomenal states and properties (which we are taking for granted) makes Physicalism false.

Examining this question will lead us to explore some other questions concerning phenomenal consciousness, among those mentioned above: we will examine some epistemological and semantic questions that will be relevant. We will see what these are, and why they are relevant, in due course.

2.2. Epistemic Arguments against Physicalism

The most powerful arguments against physicalism concerning consciousness (and perhaps, the most powerful arguments against physicalism *tout court*) are a series of arguments that we will call *epistemic arguments*.[29] These arguments are characterised by their use of epistemic premises in order to draw metaphysical consequences. In particular, the metaphysical conclusion is that phenomenal properties are not physical (in the broad sense), or more generally, that physicalism is false. Different versions of the arguments use slightly different epistemic premises. We will examine these different versions in a great detail later in this chapter and the next one, but just to get

[29] Here I follow the excellent exposition in Chalmers (2002c: 197-203).

a flavour of the arguments, let me introduce some of the main versions (in a very rough formulation). The so-called *conceivability argument*, for instance, infers, from the fact (if it is a fact) that we can *conceive* of beings physically identical to us without any phenomenal consciousness, that phenomenal consciousness does not supervene on the physical. The *explanatory argument* infers, from the alleged fact that we cannot offer an *explanation* of phenomenal consciousness in physical terms, that phenomenal consciousness itself is not physical. The *knowledge argument* infers the falsity of physicalism from the (again) alleged fact that we are not able to *deduce* truths about phenomenal consciousness from a complete description of the world in physical terms.

As we see, these arguments proceed by stating some epistemic facts about consciousness, and then inferring some metaphysical conclusion from them. These epistemic premises are about what we can conceive, or what we can explain, or what we can deduce. From these premises, the advocates of epistemic arguments infer that phenomenal consciousness is not physical, and that therefore physicalism is false. We can say that, in general, these arguments try to establish the existence of an *epistemic gap* between the physical and the phenomenal (e.g., that we can conceive of the separation of the physical and the phenomenal, or that we cannot deduce phenomenal truths from physical truths), and from this they infer the existence of an *ontological gap* between the physical and the phenomenal (i.e., phenomenal properties are not physical properties).

Our main question will be whether this kind of inference is warranted, that is, whether it is correct to infer the existence of an ontological gap of that sort from the existence of an epistemic gap of those sorts. This question is highly relevant, as we have seen, for the metaphysical issues involving consciousness, since if these epistemic arguments are successful, that means that consciousness is not physical. It is also relevant, of course, to the general issue of physicalism, which is of interest to many philosophers outside philosophy of mind. But it is also interesting for reasons that are independent from both consciousness and physicalism. For what we are examining here is whether a certain kind of epistemic argument is warranted or not. We want to know whether we are justified in extracting metaphysical consequences from epistemic premises. And this kind of inference is present in many other areas of philosophy. Indeed, the question that will concern us here is a question that pertains to the area of *methodology of metaphysics*, which considers how we can come to know

facts about metaphysics, and how we can go to answer metaphysical questions, among other questions. These issues have obvious significance. We will not explore them directly, though, but we will be deeply engaged with an applied issue in the methodology of metaphysics: can we find out whether a certain claim in ontology, namely, physicalism, is false, just by finding out what we can conceive, or explain, or deduce?

As we will see, I will be defending a negative answer to that question. That is, I will be arguing that we *cannot* find out whether physicalism is false just by examining what we can conceive, or explain, or deduce. So I will be arguing that a particular proposal about how to answer metaphysical questions about consciousness is wrong. My main aim is to criticise this kind of proposal, rather than suggesting a new, improved one. In any case, I hope that what I have to say in what follows will give some indications about how to build up a more successful methodology for metaphysical research.

The plan is as follows. In the remainder of this chapter, I will introduce a central epistemic argument against physicalism, namely, the conceivability argument, and we will examine different formulations of it in great detail. We will arrive at a very sophisticated version of the conceivability argument, namely, the so-called two-dimensional argument against physicalism, which in my view poses one the strongest challenges to physicalism in the contemporary literature. In the following chapter, I will turn to two arguments which can be seen as precursors of the conceivability argument, namely, Saul Kripke's modal argument and Stephen White's property dualism argument, and I will compare them with the two-dimensional argument. Finally, we will briefly examine Frank Jackson's knowledge argument. This will complete our presentation of the epistemic arguments against physicalism. We will then be ready to examine the main problems with these arguments.

2.3. The Conceivability Argument

The conceivability argument has received many different formulations. It could be argued that the label 'conceivability argument' does not refer to a single argument but rather to a family of arguments that are closely related, which can be labelled *conceivability arguments*. The main idea behind all conceivability arguments can be put like this. The arguments begin by stating that we can *conceive of* the separation of

the physical and the phenomenal. That is, we can conceive of the physical without the phenomenal, or the phenomenal without the physical. From this, it is inferred that the phenomenal and the physical are distinct properties.

How can this inference work? Can we infer that two entities are distinct just by being able to conceive that they are separated? This seems to be a bad sort of inference: there are many pairs of entities that we can conceive as being distinct (e.g. the morning star and the evening star) which are not really distinct.

The appeal of contemporary conceivability arguments largely relies on how they surmount this difficulty. What they do is to go, first, from matters about what is *conceivable* (epistemic matters), to matters about what is *possible* (*modal* matters). And then, they go from matters about what is possible to matters about what the *actual* world is like (*ontological* matters).

To see how this pattern of inference is supposed to work, we will first examine a version of the conceivability argument which is perhaps one of its most intuitive formulations. This is the so-called *Zombie Argument*, and has been prominently defended by David Chalmers, among others.[30]

What are zombies, in the philosophical sense? Zombies are individuals physically identical to us, molecule per molecule. They are perfect physical duplicates of us (and therefore, functional duplicates as well). They behave in exactly the same ways as we do. But there is a big difference: they are not phenomenally conscious at all. Everything is "dark inside" for them: there is nothing it is like to be a zombie. When I am in pain, for instance, my zombie is exactly in the same physical state, but she is not feeling any pain.

According to the Zombie Argument, these zombies are conceivable, that is, we can conceive of them without entertaining any contradiction. The idea of such a zombie is not incoherent. We may have good reasons to think that zombies do not exist in the actual world, but this does not mean that they are inconceivable. The argument then infers, from the conceivability of zombies, that they are metaphysically possible, that is, there is a possible world where there are zombies. The argument can be put as follows:

The Zombie Argument

[30] A classical source is Kirk (1974). See also Chalmers (1996) and (2003).

Z1: It is conceivable that there are zombies.

Z2: If it is conceivable that there are zombies, then it is metaphysically possible that there are zombies.

Z3: If it is metaphysically possible that there are zombies, then phenomenal consciousness is not physical.

_____

Z4: Phenomenal consciousness is not physical.

In what follows, we will see in great detail what the conceivability of zombies exactly amounts to, and the motivation for the crucial second premise (the inference from conceivability to possibility). But before doing that, let me add some comments of clarification regarding premise Z3.

As we have seen earlier, the conclusion of the argument, Z4, poses a problem for physicalism when we understand 'physical' in the broad sense, that is, 'physical_b' as characterized above. That is, the conclusion, so understood, amounts to the claim that consciousness (i.e. phenomenal properties) do not supervene *globally* on the physical. In particular, Z4, so understood, entails that there is a minimal physical duplicate of the actual world that is not a physical duplicate concerning phenomenal properties.

If we understand Z4 in this way (that is, as a denial of Physicalism), and we understand Z3 accordingly, then Z3 is false. For the possibility of zombies, such as we have characterized them, does not entail that phenomenal properties are not physical_b. That is, the possibility of zombies, by itself, does not entail that Physicalism is false. This is so because Physicalism has been characterised in terms of *global* supervenience: what is at stake is whether any possible world that is a minimal physical duplicate of @ is a duplicate simpliciter or not. Therefore, what is at stake is whether there is a minimal physical duplicate that lacks phenomenal consciousness. The possibility of individual zombies is not, per se, a problem for physicalism, because perhaps all those zombies are located in possible worlds that are very different, in physical respects, from the actual world. Maybe all those zombies are situated in environments wildly different from that of the actual world. And this would be perfectly compatible with the thesis of Physicalism, in the minimal characterization that we have provided, since Physicalism entails merely that any

possible world that is identical to @ in all physical respects will be identical to @ in *all* respects.[31]

Therefore, what should worry physicalists in not the possibility of individual zombies, but rather the possibility of a world physically identical to @ but different from it in phenomenal respects. For instance, let's consider a possible world physically identical to @ (in particular, a *minimal* physical duplicate) where all individuals are zombies (no organism is phenomenally conscious at all). If this is indeed a possibility, it is clear that Physicalism is false. Chalmers has called these worlds 'zombie worlds' (2003: 105). So we can formulate a more adequate version of the conceivability argument, in terms of zombie worlds:

The Zombie Argument (Take 2)

ZW1: Zombie worlds are conceivable.

ZW2: If zombie worlds are conceivable, then zombie worlds are metaphysically possible.

ZW3: If zombie worlds are metaphysically possible, then phenomenal properties are not physical_b.

———————

ZW4: Phenomenal properties are not physical_b.

As we can see, this new formulation of the argument does not suffer from the problem the previous one did. The third premise, ZW3, is now correct: the (metaphysical) possibility of zombie worlds is clearly a problem for the global supervenience of phenomenal properties. In other words, the possibility of zombie worlds would entail that there are minimal physical duplicates of @ that are not duplicates *simpliciter* and, therefore, that Physicalism is false.

We could also build an argument against physicalism in terms of individuals with an inverted spectrum, rather than individuals that are not conscious at all. Inverted-spectrum individuals are such that they are physically identical to us molecule by molecule, but their experiences are systematically inverted from those

[31] As we saw in the previous chapter, the notion of supervenience that best captures the intuitive idea of physicalism is global supervenience rather than local supervenience. Here we are just rehearsing the argument that we offered then: local supervenience seems too strong, since it is committed to the claim that all individuals physically identical to us will be phenomenally identical too. This view does not allow any phenomenal variation due to physical differences in the environment, which is something that physicalism should allow.

that we have. If, for instance, I am having a sensation of seeing something blue, my corresponding inverted-spectrum twin is having the sensation of seeing something red, and so on. These beings are physically identical, but phenomenally different. So let 'inverted-spectrum worlds' be possible worlds that are minimal physical duplicates, but where conscious beings have their experiences inverted in such sense. Then, this version of the conceivability argument would go as follows:

The Inverted Spectrum Argument

IS1: Inverted-spectrum worlds are conceivable.

IS2: If inverted-spectrum worlds are conceivable, inverted-spectrum worlds are metaphysically possible.

IS3: If inverted-spectrum worlds are metaphysically possible, Physicalism is false.

———————

IS4: Physicalism is false.

Indeed, we can see that any minimal physical duplicate where any fact concerning phenomenal consciousness is different will pose a problem for Physicalism. For instance, let's assume that at time *t* I am having a toothache (in the actual world). Then, the possibility of a minimal physical duplicate where all the phenomenal facts are identical with the sole exception of my corresponding counterpart not feeling a toothache at time *t* would also falsify physicalism.

Hence, we can build a new version of the conceivability argument in general terms, as follows. Let P be a complete description of the world in physical_n terms, that is, in terms of the entities and properties of the sort posited by current physics. P will include expressions to refer to each fundamental physical entity of the world, and will say what fundamental physical properties those fundamental physical entities have. Let Q be an arbitrary phenomenal truth about the actual world: for instance, the truth that I am having a toothache at *t*, or that I am phenomenally conscious, or even a complete phenomenal description of the world (that is, a description of all the phenomenal properties that are instantiated in the actual world, either at a given time, or during a certain interval of time).

Physicalism, as characterized above, entails that for any possible world where P is true, Q has to be true as well. More precisely, any *minimal* physical duplicate

where P is true, has to satisfy Q as well. (In other words: any possible world which satisfies P and a 'that's all' clause has to satisfy Q as well. Let T be this 'that's all' clause, and let PT stand for the conjunction P&T. Then, Physicalism entails that in any possible world where PT is true, Q is true too. In what follows, we will understand P as including the clause T.) That is, the possibility of a world where P holds but Q does not would falsify Physicalism. So we can formulate the conceivability argument as follows:

The Conceivability Argument

CA1: P&~Q is conceivable.

CA2: If P&~Q is conceivable, then P&~Q is metaphysically possible.

CA3: If P&~Q is metaphysically possible, physicalism is false.

———————

CA4: Physicalism is false.

This is the formulation of the conceivability argument that we will be mostly concerned with, since it is clearly more general. The zombie argument and the inverted spectrum argument can be seen as instances of this more general argument. They represent two of the many ways in which P&~Q could be satisfied: by means of a zombie-world, or by means of an inverted-spectrum world. Indeed, any possible world physically identical to @ where any phenomenal truth Q did not hold would be a possible world where P&~Q would hold (for the corresponding Q).

Thus, the crucial questions are the following: first, whether P&~Q can be conceived without contradiction, and second, whether the conceivability of P&~Q would entail its possibility. As we will see, it is a difficult enterprise to define a single notion of 'conceivability' that makes premises CA1 and CA2 both true. However, Chalmers has argued that if we reason carefully, we can move from conceivability to possibility.[32] In particular, he has distinguished several notions of conceivability and possibility, and has singled out a notion of conceivability such that, he argues, P&~Q is conceivable in that sense and its conceivability entails its possibility, in the sense of 'possibility' that makes CA3 true. This would be a very important project, if successful. In the rest of this chapter we will explore these different notions of

[32] See Chalmers (1996), (2002a), (2003) and (forthcoming).

conceivability and possibility, and Chalmers' proposal about how to best understand the conceivability argument. In the rest of this thesis, we will be largely concerned with the critical assessment of the conceivability argument, understood along those lines.

2.4. Notions of Conceivability

In this section I aim to explain the notion of conceivability that is invoked in Chalmers' conceivability argument, and which is such that, according to him, a sentence being conceivable in that sense entails its being possible. We will introduce this notion by comparing it with other related notions of conceivability. Chalmers' strategy is to examine these diverse sorts of conceivability, and the different problems attached to the claims that they entail possibility. After examining those problems, Chalmers concludes that there is a notion of conceivability which can overcome them, so that the claim that conceivability in this sense entails possibility is tenable.[33]

2.4.1. Prima facie vs. ideal conceivability

The notion of conceivability that is crucial to Chalmers' conceivability argument can be characterised, on a first approximation, as follows: a sentence S is conceivable when S is not incoherent or contradictory. By this, we mean *logically* or *conceptually* incoherent. That is, S is conceivable when there is no contradiction in it that we can discover merely by conceptual means. In other words, S is conceivable when it cannot be ruled out *a priori*, that is, when it cannot be ruled out independently of experience. Hence, S is conceivable when S is not a priori false.

This is the notion of conceivability that we will focus on. One problem that has been raised with respect to the claim that conceivability in that sense entails possibility is the following. Take a correct but non-trivial mathematical claim M. There can easily be subjects who are not able to realise, by a priori reasoning, that M is true. That is, these subjects would not realise that the negation of M (~M) is false. Therefore, ~M would be conceivable for them, since they cannot rule it out a priori. But, clearly, ~M is not possible, since M is a necessary truth.[34]

[33] This strategy is especially developed in Chalmers (2002a), from which I draw in what follows.
[34] We are assuming that mathematical truths are necessary truths.

In order to solve this kind of difficulty, Chalmers distinguishes between *prima facie* conceivability and *ideal* conceivability. A sentence S is prima facie conceivable for a subject when S passes the corresponding test of conceivability for this subject, on first appearances. In this way, this distinction assumes that we have fixed some criterion for what conceivability amounts to. Once this criterion is fixed, we can assess whether a certain sentence S satisfies such a criterion, on first appearances, or not. Here we are going to focus on the test of conceivability already mentioned, that is, that S is not ruled out a priori. Therefore, we can characterise prima facie conceivability as follows:

> S is *prima facie conceivable* for a subject H if and only if H cannot rule S out a priori after some reflection.

As we see, the notion of prima facie conceivability is always relative to some subject. It is perfectly possible that, for a certain sentence S, S is prima facie conceivable for some subjects but not for others.

It seems clear that this notion of conceivability does not entail possibility, since there are sentences that are clearly impossible but are nonetheless prima facie conceivable for some subjects. As we have seen, some mathematical falsehoods are examples of this: they can be prima facie conceivable for some subjects, but not for others, and in the last instance, those sentences are not true in any possible world.

Therefore, any notion of conceivability that could serve as a guide to possibility would have to be more restricted than prima facie conceivability. Chalmers introduces the notion of *ideal* conceivability for that purpose: a sentence is ideally conceivable when it passes the corresponding test of conceivability, even for ideal subjects, that is, after ideal reflection. When we apply our criterion of conceivability (i.e. not being ruled out a priori) into this schema, the characterization of ideal conceivability that follows is like this:

> S is *ideally conceivable* if and only if an *ideal* subject cannot rule S out a priori.

As we have seen, there are many sentences that we cannot rule out a priori, because they are too complex or too long to process, but there might be subjects with stronger

reasoning abilities who are indeed able to rule them out a priori. Therefore, what we need is a test for conceivability in terms of ideal subjects, that is, subjects who are free from our contingent cognitive limitations. Then, we can say that a sentence is ideally conceivable when even an ideal subject (free from our contingent cognitive limitations) could not rule it out a priori.

What is an ideal subject exactly? It is hard to offer a substantial definition of what an ideal subject is, and in effect Chalmers does not attempt to provide one. However, I think that it can be useful to see ideal subjects as subjects who can find out any a priori truth that there is. This is still not very precise, but it will suffice for our purposes. In this way, a sentence will be ideally conceivable when it is not ruled out a priori, not even by an ideal subject who could find out all the priori truths.

Alternatively, if we have doubts about the coherence of the notion of an ideal subject, we can understand ideal conceivability not in terms of ideal subjects but in terms of ideal reflection. Ideal reflection is opposed to prima facie reflection, and it can be characterised as reflection which cannot be defeated by further reflection: [35]

> S is *ideally conceivable* if and only S cannot be ruled out a priori after *ideal* reflection.

In any case, Chalmers does not provide a substantial definition of what ideal reflection is either. His strategy is rather to trust in our intuitive grasp of these notions, and to take them as primitive rational notions. His aim is to explore the connection between these intuitive rational notions (i.e. ideal, undefeated a priori reflection) and the modal notion of possibility, to see whether there is any tenable link between the two.[36]

We can conclude that ideal conceivability is a better guide to possibility than prima facie conceivability. Therefore, given a certain inference from conceivability to possibility, the closer the notion of conceivability used is to that of ideal conceivability, the more chances of being successful such inference will have. We cannot be sure that a given sentence is conceivable after ideal reflection, since we cannot be sure that our reflection cannot be defeated by further reflection. But we can

[35] That is, reflection which cannot be defeated by further reflection by anyone: this notion is no longer subject-relative.
[36] See Chalmers (2002a: 147-49).

at least try to make the case as strong as we can, by subjecting the sentence that is supposed to be conceivable to serious scrutiny. If we cannot find any contradiction, after serious reflection, then we can make a good case for that sentence's being ideally conceivable.

2.4.2. Negative vs. positive conceivability

The notions of conceivability we have been discussing so far are all instances of what Chalmers calls 'negative conceivability'. A sentence is negatively conceivable, in general, when S cannot be ruled out. There are different ways of spelling this out: S cannot be ruled out for all we know, or S cannot be ruled out given the best scientific theories, or S cannot be ruled out by a priori methods. We have been focusing on the latter notion, since we are interested here in the connection between *a priori* notions of conceivability and the notion of possibility. Therefore, we can focus on the following formulation of negative conceivability:

> S is (*ideally) negatively conceivable* if and only if we cannot rule it out a priori, after ideal reflection.[37]

This is the most promising notion of conceivability that we have got so far. However, there are still some problems for the claim that conceivability, so understood, entails possibility. One problem is the following. Let's consider a case of a non-trivial mathematical claim, which has not been proved nor refuted yet. For instance, let's take the case of Goldbach's conjecture (GC, henceforth). This conjecture seems likely to be true, so we will assume, for the sake of the argument, that GC is true. However, no conclusive proof has been offered yet. Then, the negation of Goldbach's conjecture (~GC) is prima facie negatively conceivable for us. The question is, is it ideally conceivable? Well, it depends on whether there is, at the end of the day, an a priori

[37] As we can see, this notion of negative conceivability is a version of *ideal* conceivability, in terms of what can be ruled out a priori after ideal reflection (or for ideal subjects). We can also formulate a notion of *prima facie* negative conceivability, in terms of what can be ruled out for a given subject:

S is *prima facie negatively conceivable* for a subject H if and only if H cannot rule it out a priori after some reflection

As we have seen earlier, ideal conceivability is a better guide to possibility than prima facie conceivability, so we will largely focus on *ideal* rather than prima facie negative conceivability.

proof for the truth of GC or not. It could be argued that there is the possibility that there was no a priori proof for GC: not only have we not found it yet, but in addition there is no conclusive a priori proof for such a conjecture in logical space, not even for an ideal subject. If this were the case, then ~GC would be not only prima facie but also ideally negatively conceivable, since it could not be ruled out a priori after ideal reflection. In any case, in this scenario ~GC will not be possible, because we have assumed that GC is true, and therefore, since it is a mathematical truth, it is necessarily true. So if GC were true but no a priori proof were available, not even for an ideal subject, then this would pose a counterexample to the claim that (ideal) negative conceivability entails possibility. [38]

Chalmers agrees that this kind of example could pose a problem, although he believes that it is not completely clear that such scenarios are really coherent: he has doubts about the possibility of a mathematical truth for which there is no a priori proof, not even after ideal reflection: "In any given case, one can argue that either (1) the statements in question are knowable under some idealization of rational reasoning, or (2) that the statements are not determinately true or false" (Chalmers (2002a: 180)). Nonetheless, he recognises that there is more to say about this issue, and that he cannot rule out in a conclusive manner that there could be cases where negative conceivability fails to entail possibility. In order to solve this problem, he distinguishes negative conceivability from what he calls *positive conceivability*, and he claims that positive conceivability is a stronger guide to possibility.[39]

> S is *positively conceivable* if and only if one can coherently imagine a situation in which S is the case.[40]

The main idea here is that a sentence S is conceivable in this sense when we are able to form a positive conception of it, that is, when we can imagine a detailed *scenario* such that it entails that S is true. What does it mean to say that an imagined scenario entails that S is true? Chalmers (2002a) explores this question in great detail, but it will suffice for our purposes to give a brief introduction here.

---

[38] The case of Goldbach's conjecture and some other similar cases are discussed in Chalmers (2002a: 160-1 and 180).
[39] In particular, Chalmers says that "due to its added strength, [ideal positive conceivability] is in a slightly better position to be a perfect guide than [ideal negative conceivability]" (2002a: 161).
[40] Chalmers (forthcoming: 3).

First, it should be noted that scenarios are characterised in epistemic terms: we should not presuppose that when someone is imagining a scenario, there is always a corresponding metaphysically possible world. The question at issue in this debate is to define a purely epistemic notion of conceivability, and to assess whether it entails possibility or not. So, in order to avoid begging the question, we should not characterise imagined scenarios or situations as metaphysically possible worlds: we can instead see scenarios as *complete descriptions* of ways the world might be.

Secondly, these complete descriptions or scenarios should be such that they contain no a priori incoherence. As Chalmers says, "a situation is coherently imagined when it is possible to fill in arbitrary details in the imagined situation so that no contradiction reveals itself" (2002a: 153). That is, after we fill in all the gaps in the imagined scenario, no contradiction can be found.

Thirdly, we have said that S is positively conceivable when we can coherently imagine a scenario where S is true. And by this we mean a scenario such that its description entails that S is true. What does this mean? According to Chalmers, the central sort of entailment here is that of *a priori entailment*. S is true in an imagined scenario if and only if the corresponding description entails S a priori. That is, if we can imagine that scenario obtaining without S obtaining, then the scenario does not entail S a priori, and therefore S would not be true at that scenario.

Fourthly, we can distinguish two notions of positive conceivability, according to whether it is prima facie or ideal conceivability:

> S is *prima facie positively conceivable* for H if and only if H can coherently imagine a scenario where S is true (i.e. H takes it that the description of that scenario a priori entails S).

As we saw earlier, prima facie conceivability is relative to some subject: a sentence will be prima facie positively conceivable for a subject when this subject can imagine a (prima facie) coherent scenario such that it (prima facie) entails a priori the truth of such a sentence. This prima facie conceivability can be defeated at least in two ways. Perhaps after filling in all the arbitrary details of such a scenario, we realise that there is some a priori contradiction, so that the scenario is no longer coherently imaginable. Or perhaps after reflection we realize that such scenario does not entail a priori the truth of S after all: maybe we misdescribed the situation and the sentence that is really

entailed by the scenario is a different one. One example of this possible mistake is the following. It could seem to us that it is easy to imagine a situation that entails the truth of the sentence 'Goldbach's conjecture is false': we can just imagine a situation where important mathematicians announce that Goldbach's conjecture has been conclusively falsified. But a bit of reflection will show that this scenario does not really entail a priori that GC is false. For it is coherent to imagine a situation in which mathematicians announce that GC is false but it is true nonetheless. A sentence that would clearly be entailed a priori by such scenario is, for instance, 'mathematicians have announced that GC is false'. But this does not entail a priori that GC is false. Therefore, this is not really a scenario where GC is false.

So some cases of prima facie positive conceivability could be defeated after further reflection. In order to account for these cases, Chalmers also introduces a notion of ideal positive conceivability:

> S is *ideally positively conceivable* if and only if we can coherently imagine a scenario where S is true (i.e. S is a priori entailed by the description of that scenario) after ideal reflection.

As Chalmers says, "when S is ideally positively conceivable, it must be possible in principle to flesh out any missing details of an imagined situation that verifies S such that the details are imagined clearly and distinctly, and such that no contradiction is revealed. It must also be the case that rational reflection on the imagined situation will not undermine the interpretation of the imagined situation as one in which S is the case" (2002a: 153).

The notion of ideal positive conceivability is the most promising one so far, as a guide to possibility. As we have seen, ideal notions of conceivability are always stronger guides to possibility than their corresponding prima facie ones. It could also be argued that there are cases of negative conceivability that do not entail positive conceivability, and therefore they do not entail possibility either.[41] If this is truly the

[41] We are assuming here that possibility entails positive conceivability, and that positive conceivability entails negative conceivability. What is at issue is whether the reverse implications hold or not. It could seem controversial to claim that if S is possible, then S is positively conceivable, but we should remember that we are focusing on the notion of ideal positive conceivability. We can motivate the claim that positive conceivability entails negative conceivability as follows. We cannot have cases of (ideal) positive conceivability without (ideal) negative conceivability: if a sentence is not negatively

case, then positive conceivability will be a stronger guide to possibility than negative conceivability.

One alleged case of failure of inference between negative and positive conceivability is that of the Goldbach's conjecture (GC). As we have seen above, it could be argued that the following hypothesis is coherent: (a) GC is true, and (b) there is no a priori proof for GC, not even for an ideal subject. Then, ~GC is (ideally) negatively conceivable, but not possible. Crucially, ~GC would not be (ideally) *positively conceivable* though: it just does not seem possible to imagine a coherent scenario which a priori entails ~GC. As we said before, we could easily imagine a scenario where mathematicians announce that they have found a proof of ~GC, but this does not really count as a scenario where ~GC is true, because such scenario does not really entail a priori ~GC. In order to positively conceive of ~GC, we should find a scenario that entails a priori that GC is false, that is, we would need a proof that GC is false. And by hypothesis, GC is true. So we could not really imagine a scenario where ~GC is true.

If this argument is accepted, then it follows that negative conceivability does not always imply positive conceivability. But this kind of case does not pose any obstacle to the claim that positive conceivability implies possibility: in our putative counterexample, possibility fails but positive conceivability fails as well.

As I have noted earlier, Chalmers is not sure that this sort of case poses an insoluble problem for the link between negative and positive conceivability, because it is not clear that (a) and (b) above are really consistent. However, there might be similar cases that are more difficult to solve. Therefore, Chalmers concludes that positive conceivability is at least as strong a guide to possibility as negative conceivability, and perhaps positive conceivability is a stronger guide than negative conceivability. So we will have a stronger case for the possibility of a certain sentence if we can show that such a sentence is not only negatively but also positively conceivable.

2.4.3 Primary vs. secondary conceivability and possibility

We have examined several notions of conceivability, and we have concluded that the most promising guide to possibility, among the notions of (a priori) conceivability

---

conceivable, then it can be ruled out a priori and therefore there could be no *coherent* scenarios where it were true, that is, the sentence would not be positively conceivable.

discussed so far, is that of ideal positive conceivability. However, there are still important challenges to the claim that such notion of conceivability entails possibility. The main challenge to this claim comes from the familiar cases of a posteriori necessities discussed by Saul Kripke in *Naming and Necessity*, such as the following:

(1) Water is H_2O

(2) Heat is molecular motion

Kripke (1980) famously argued that these sentences are necessary truths, that is, they are true in every possible world.[42] He argues that natural kind terms such as 'water' and 'heat' are *rigid designators*, that is, terms that refer to the same entity/kind in every possible world. If this is correct, then it follows that identity statements such as (1) and (2) are true in every possible world. For it is widely accepted that the right-hand side terms, that is, 'H_2O' and 'molecular motion' are rigid designators too. Now, if, say, sentence (1) is true in the actual world, it means that 'water' and 'H_2O' are co-referential in the actual world. And since each term is a rigid designator, this means that each of them will have the same referent in every possible world. Therefore, both terms will be co-referential in every possible world. That is to say, (1) will be true in every possible world. Likewise with respect to sentence (2).[43]

So we can conclude that those sentences are metaphysically necessary. But let's now consider their negations:

(~1) Water is not H_2O

(~2) Heat is not molecular motion

[42] It could be argued that these sentences cannot be true in possible worlds where the main expressions do not refer to anything, that is, in possible worlds where 'water' or 'heat' do not have a referent. If so, then we can say at least that (1) and (2) are true in every possible world in which their terms have a referent. Alternatively, we can say that the sentences that are really necessary are not (1) and (2) but rather (1'): 'If water exists, water is H_2O' and (2'): 'If heat exists, heat is molecular motion'. We can ignore this complication in what follows.

[43] As we can see, the inference from the terms being rigid designators to the corresponding identity statements being necessary is rather trivial. The crucial part of Kripke's argument relies on his defence of the claim that the terms 'water' and 'heat' are in effect rigid designators. He motivates this by examining our intuitions concerning how we would use such natural kind terms with respect to counterfactual possibilities of different sorts. He argues that, for instance, no matter what appearance H_2O could have in a given possible world, we would still apply the term 'water' to that stuff. See Kripke (1980: Lectures I and II) for further discussion.

Are these sentences conceivable? It seems clear that they are, at least, prima facie negatively conceivable, since it seems clear that we cannot rule them out a priori. That is, they do not involve any obvious a priori contradiction. But it also seems that no further rational reflection could reveal any a priori contradiction. How could (~1) contain any hidden a priori contradiction? What (~1) says seems perfectly coherent, even if false. That is, before we learn chemistry, we think that (~1) is perfectly plausible, an open epistemic possibility. So we have good grounds to assert that (~1) is ideally negatively conceivable, and likewise with respect to (~2).

However, if Kripke is right, then (~1) and (~2) are not metaphysically possible, since (1) and (2) are true in every possible world. So (~1) and (~2) seem to be important counterexamples to the claim that (ideal negative) conceivability entails possibility.

Chalmers' response to this problem constitutes one of the crucial steps in his defence of the conceivability-possibility link. Let me explain how this response goes, in some detail.

The intuitive idea is this. When we conceive of sentence (1) being false (or sentence (~1) being true), what we conceive of is basically a world where the stuff that looks like water is not H_2O, or where H_2O does not have the superficial properties of water. For instance, we can focus on Twin Earth (which was introduced in the previous chapter), which is a world of that sort: the stuff that looks like water there is XYZ, whereas H_2O has very different superficial properties, such as being black and tarry. So when we conceive of (~1) being true, what we conceive is that a scenario like this is the case. Now, it is crucial to notice that these scenarios seem to be *metaphysically possible* indeed. Nothing prevents the metaphysical possibility of a world where the stuff that fills river and lakes and so on (the "watery stuff") has a different chemical composition. If so, nothing prevents the possibility of a world where XYZ has the same behaviour and appearance as H_2O has in the actual world. This may seem a far-fetched possibility, but it is a possibility nonetheless.

Therefore, Chalmers thinks that Kripke's familiar examples still point out some connection between conceivability and possibility: for each one of these cases of *conceivable* sentences, there is always some *metaphysical possibility* accompanying them. This is the suggestion that he wants to defend.

First, Chalmers introduces a distinction between what he calls primary conceivability and secondary conceivability. *Primary conceivability* has to do with

purely conceptual, a priori matters, so this is the notion of conceivability that we are interested in here. Negative primary conceivability is a matter of the truth of a sentence not being ruled out a priori, and positive primary conceivability is a matter of our ability to coherently imagine a scenario that a priori entails a sentence. On the contrary, *secondary conceivability* appeals to our empirical knowledge of the actual world, in order to find out what is conceivable in this sense or not. For example, once we know that water is indeed identical to H_2O, it turns out that (~1) is not secondarily conceivable. Likewise, once we know that heat is identical to molecular motion, (~2) is no longer secondarily conceivable.[44]

So one response to the Kripkean cases above would be to say that whereas primary conceivability may fail as a guide to possibility, secondary conceivability might still very well be a good guide, for all those counterexamples say. Chalmers' response goes further than this, though. His aim is to find an *a priori* notion of conceivability that can serve as a guide to possibility, and for this reason secondary conceivability is not enough for his purposes. He wants to explore to what extent primary conceivability is a guide to possibility. And he suggests that it does entail possibility, once we are very careful about what kind of possibility is at issue here.

As we have seen, there is a sense in which the (primary) conceivability of (~1) is accompanied by a genuine metaphysical possibility, such as Twin Earth. There is a sense in which sentence (~1) is true at Twin Earth. There is a way of evaluating (~1) such that it turns out to be true at Twin Earth. What way is this? According to Chalmers, when we consider Twin Earth *as actual*, (~1) turns out to be true. What does this mean? Well, when we consider a possible world as actual, we take it to be the world we are in, the world where the references of our terms are fixed. That is, we imagine that that is the way the actual world is.

So, let's imagine that Twin Earth were our actual world. If that was the way our actual world was, then the watery stuff of our acquaintance would be XYZ, not H_2O. And, what would 'water' refer to, if Twin Earth were our actual world? It seems clear that 'water' would refer to XYZ, because 'water' refers to whatever happens to be the watery stuff of our acquaintance in a given possible world considered as actual. Let me elaborate this a little.

[44] In what follows, we will focus mainly on primary conceivability, so I will not say more about what secondarily conceivability amounts to. See further discussion in Chalmers (2002a), and also in sections 2.4.3 and 2.5 below, where I explain what *secondary possibility* is, which is closely related to secondary conceivability.

There are different theories of reference that would entail that 'water' would refer to XYZ in Twin Earth considered as actual. For instance, the so-called *descriptivist theory of reference* clearly has that consequence. This theory claims that the reference of 'water' (and other natural kind terms) is fixed by means of a description associated with the term.[45] The description would be something along these lines:

> $WATER_d$: The colourless, odourless stuff of our acquaintance that fills rivers and lakes, falls from the sky, is called 'water' by experts, and so on.

The descriptivist theory of reference claims that 'water' refers to whatever satisfies the description in a given world considered as actual.[46] So, if Twin Earth turned out to be the actual world, then XYZ would satisfy that description, and 'water' would refer to XYZ.

Therefore, holding a descriptivist theory of reference is one way of getting to that conclusion. But as Chalmers notices, it is not the only way. We do not have to be committed to a descriptivist theory of reference, in order to claim that the reference of 'water' in Twin Earth considered as actual is XYZ. Chalmers believes that the minimal commitment of such a claim is just the following: the basic idea is that competent users of a term have the ability to find out what the referent for a certain term is, given enough information about a scenario. According to him, competent users have intuitions about what a certain term would refer to, if the actual world turned out to be this or that way. Of course, if the reference of a term such as 'water' is fixed by a description that is a priori associated with the term, then it is easy to see

---

[45] Typically, descriptivist theories of reference hold that these descriptions are associated *a priori* with the corresponding terms by the competent users of the terms.

[46] Notice that according to this version of the descriptivist theory, the description $WATER_d$ fixes the reference of 'water', but this version is not committed to the term and the description being synonymous. In Kripke's words, the description can be reference-fixing without being meaning-giving. In particular, the description above could fix the reference of 'water', but it could not give the meaning because such description is non-rigid, and as we have said, 'water' is a rigid designator, so only a rigid description could give its meaning. Nonetheless, we can formulate a *rigidified* version of the description, which could arguably be meaning-giving as well as reference-fixing. The description could go as follows:

> $WATER_d$*: The *actual* colourless, odourless stuff of our acquaintance that fills rivers and lakes, falls from the sky, is called 'water' by experts, and so on.

This description will refer, in any possible world, to the stuff that plays the water role in the actual world, that is, H_2O, just as 'water' does. See, among others, Braun (2006) and Stanley (1997) for further discussion.

that we are going to be able to find out what the referent of 'water' is, given enough information about a possible world. For we just have to find out what entity satisfies the associated description, and that entity will be the referent of the term, in that possible world considered as actual. But even if there are terms that are not associated with descriptions, we could still be able to find out their referents, given a description of a possible world considered as actual. According to Chalmers, our intuitions about what a term would refer to in different possible worlds considered as actual are more basic than the descriptions that we might associate with the term. For some terms, our intuitions can yield an easily statable description, posing necessary and sufficient conditions for an entity to be the referent of the term. But for other terms, our intuitions are not so neat and tidy, and they do not easily provide us with a simple set of necessary and sufficient conditions. This is no obstacle, though, to the idea that we could find out the referents of the term in different possible worlds considered as actual: what we need for this is just a dispositional ability to identify the referents of a given term in different scenarios. And we could very well have this dispositional ability with respect to a certain term, without having to associate this term with a reference-fixing description.[47]

Actually, as Chalmers and Jackson have argued, this basic idea, i.e. that we are able to evaluate different possible worlds considered as actual, is compatible with (at least some versions of) the main rival to the descriptivist theory of reference, namely, the causal theory of reference.[48] The causal theory of reference was famously suggested by Kripke (1980) and the main idea behind it is that natural kind terms (among other referring terms) get the reference they have by means of causal-historical chains that go from a speaker's use of a certain term, all the way back to the original name-dubbing of the referent. These causal-historical chains are constituted by previous speakers' uses of the term: the idea is that I use a term with a certain reference because I borrowed it from some other speaker that used the same term, who in turn borrowed it from someone else who borrowed it from someone else… and so on, until we reach the very first uses of the term, when it was first introduced to refer to such a referent. The idea here is that speakers do not have to associate descriptions with the term in order to be able to use it to successfully refer to the kind

[47] For further discussion of this point see, among others, Chalmers and Jackson (2001: 321-2) and Chalmers (2002b)
[48] See Chalmers (1996), especially pp. 56-65, and Jackson (1998a), especially pp. 37-41.

it refers to: they just have to be part of the appropriate causal chain. Chalmers and Jackson have argued that this theory of reference is not incompatible with the idea that we are able to identify the referent of a term, given a sufficiently rich description of a possible world, considered as actual. Indeed, they argue that the arguments for the causal theory of reference seem to presuppose such a view.

Let me illustrate this with a brief discussion of one such argument. This argument can be recalled from Kripke and Putnam's discussions about the descriptivist theory of reference for natural kind terms. They argue that the reference of such terms cannot be fixed by means of descriptions, since the most familiar descriptions would yield a different referent in some situations. For instance, Putnam (1975) asks us to consider Twin Earth as a remote planet in our actual universe. If that were so, then there would be some stuff, namely XYZ, which would satisfy the descriptions commonly associated with 'water', namely, being watery stuff. Nonetheless, XYZ would *not* be the referent of 'water', since it is obvious that 'water' refers to H_2O, not XYZ, in the actual universe.

The proper moral to be drawn from this thought-experiment, according to Chalmers and Jackson, is that 'water' refers to the watery stuff *of our acquaintance*, that is, the watery stuff that we have been in *causal* contact with, not any stuff that happens to be watery stuff in any remote corner of the universe. This suggests that we have intuitions about the referent of 'water' in different possible worlds considered as actual: for instance, we can see that, if Twin Earth was a remote planet in our actual universe, 'water' would still refer to H_2O, not XYZ. We can react to this information in different ways. One way would be to re-formulate the descriptivist theory of reference so that it accounts for our intuitions concerning such a scenario. The description $WATER_d$ above does account for this, since it states that the referent has to be the watery stuff of our acquaintance. However, even if we find new problems with the claim that such a description fixes the reference of 'water', the crucial idea is that we have learned about those problems by examining different possible worlds considered as actual, and by reflecting on what the term 'water' would refer to in those scenarios. Then, this ability to identify the referent in different scenarios seems to be presupposed by both parties in the debate concerning reference: it seems to be the very methodology they employ in order to argue for or against a particular view.

Hence, these considerations suggest that that the term 'water' can be evaluated in different possible worlds considered as actual, such as Twin Earth, and that it

would refer to XYZ in such a scenario. This is perfectly compatible with the claim that 'water' is a rigid designator: this means that 'water' refers to the same entity in all possible worlds considered as *counterfactual*. In order to consider a possible world as counterfactual, what we do is to keep the actual world fixed, and then we wonder what would have been the case in these other worlds that are not actual. If we were to assume that the actual world is the actual one, where 'water' actually refers to H_2O, then we would conclude that 'water' would refer to H_2O in any other counterfactual possibility. This is what our intuitions about natural kind terms tell us: by using such terms, we aim to refer to certain stuff, independently of its contingent properties. Therefore, when we talk about counterfactual possibilities, we still use the term 'water' to refer to H_2O, independently of its contingent properties.

However, when we consider a possible world as actual, we assume that it is the context where the reference of our terms is fixed. Therefore, it is plausible to say that if Twin Earth was our actual world, then 'water' would refer to XYZ. If so, then sentence (~1) would be true at Twin Earth, considered as actual:

(~1) Water is not H_2O

The idea is that, since 'water' would refer to XYZ in that possible world, then (~1) would turn out to be true. So there is a certain sense in which (~1) is *possible*: there is at least one perfectly genuine possible world *considered as actual* where (~1) is true (since Twin Earth seems to be a perfectly genuine possibility). We will call this *primary possibility*, and we will say that (~1) is primarily possible (1-possible), in this sense.

Likewise for sentence (~2), which we will recall here for convenience:

(~2) Heat is not molecular motion

There is a possible world considered as actual where this sentence is true, namely, a possible world where the phenomenon that causes sensations of heat is something other than molecular motion. If this world turned out to be the actual world, 'heat' would refer to that other phenomenon (or so our intuitions about the term 'heat' say), so (~2) would come out true.

Therefore, as we can see, sentences (~1) and (~2) are both (ideally) conceivable, and they are also possible in some sense, namely, they are 1-possible. Indeed, it seems plausible that all cases of a posteriori necessary identities of the sort discussed by Kripke, such as (1) and (2), work in this way: these sentences are true in every possible world considered as counterfactual, but not in every possible world considered as actual. Therefore, these examples are indeed compatible with some attenuated link between conceivability and possibility. In particular, they are compatible with the claim that conceivability entails primary possibility. We can formulate this claim as follows:

(CP) Ideal conceivability entails primary possibility.

We can formulate two different versions of this thesis, according to whether we understand 'conceivability' as negative or positive conceivability. In particular, these two principles will go as follows:

(CP+) Ideal positive conceivability entails primary possibility.

(CP-) Ideal negative conceivability entails primary possibility.

As we have seen, our examples (~1) and (~2) seem to fit (CP-) very well: both sentences are negatively conceivable and also primarily possible. In addition, we can also see that these examples also fit (CP+): (~1) and (~2) are perfectly compatible with the claim that positive conceivability entails primary possibility (although they do pose a problem for the claim that positive conceivability entails possibility in the standard sense).

In order to see why those cases fit (CP+), the first question to deal with is the following: are (1) and (2) positively conceivable? At the beginning of this section we argued that they seem negatively conceivable but we did not say anything about their being positively conceivable (and as we have seen, negative conceivability does not necessarily entail positive conceivability). But we are now in a position to see that (1) and (2) seem to be positively conceivable as well.

Let's focus on (~1). Could we coherently imagine a scenario whose description entails (~1) a priori? According to Chalmers (2002a) and Jackson (1998a)

the answer is yes, and they offer the example of the Twin Earth scenario to illustrate this.

Chalmers and Jackson have argued that a description of Twin Earth would entail a priori the truth of sentence (~1), by means of the following a priori inference:

(i) The colourless, odourless stuff of our acquaintance that fills rivers and lakes, falls from the sky, is called 'water' by experts and so on, is XYZ. (From the description of Twin Earth)

(ii) If something is the colourless, odourless stuff of our acquaintance that fills rivers and lakes, falls from the sky, is called 'water' by experts, and so on, then it is water. (From the definition of 'water', $WATER_d$)

(iii) Therefore, XYZ is water.

(iv) Therefore, water is not H_2O.

This is supposed to be an a priori inference from the description of Twin Earth to (~1). The crucial premise here is (ii), which is supposed to be a priori true, since it is just an application of our definition of 'water' above.[49] If this is indeed a valid a priori inference then it shows that the Twin Earth scenario entails a priori (~1), and therefore (~1) is true at that scenario. So we have good reasons to think that (~1) is positively conceivable, and in addition it seems that no further rational reflection could defeat our prima facie case for the positive conceivability of (~1). That is, we have good reasons to think that (~1) is ideally positively conceivable. (Analogous reasoning would show that (~2) is ideally positively conceivable).

Therefore, both (~1) and (~2) seem to be positively conceivable. But they do not seem to be possible in the standard sense: as Kripke argued, they are false at every possible world considered *as counterfactual*. Does this pose a problem for (CP+)? Chalmers has shown that it does not, because what is relevant with respect to (CP+) is whether (~1) and (~2) are true at some possible world considered *as actual*. And we

[49] As we saw earlier, according to the descriptivist theory of reference the description $WATER_d$ is a priori associated with the term 'water', and this is why (ii) above is a priori true. If so, then it is easy to infer (iii) a priori from (i), given that (ii) is a priori. However, according to Chalmers and Jackson, we do not have to be committed to the descriptivist theory in order to endorse the claim that (ii) is a priori true (that is, a conceptual truth). They claim that all is required is the claim that competent speakers have the disposition to identify the referents of a term in different scenarios considered as actual. (We will talk more about this issue in Ch. 4.) In any case, Jackson is more sympathetic to the descriptivist theory of reference than Chalmers: see for instance Jackson (1998b) and Chalmers (2002b).

can indeed see that there are possible worlds considered as actual where they are true. Therefore, (CP+) seems to be vindicated by these examples.

So we have arrived at two conceivability-possibility links that are both compatible with standard cases of a posteriori necessities: (CP-) and (CP+). Therefore, in the following chapters we will focus on these two claims, and we will assess whether they are ultimately correct or not.

Before closing this section, I would like to make a final comment regarding the different strength of these principles (which we can repeat here for convenience).

(CP+) Ideal positive conceivability entails primary possibility.

(CP-) Ideal negative conceivability entails primary possibility.

We can assume that primary possibility entails both negative and positive conceivability.[50] If so, then (CP-) will be committed to the controversial claim that negative conceivability entails positive conceivability.[51] As we have discussed above, there might be some counterexamples to this claim, although it is not clear whether they ultimately work. On the other hand, (CP+) is committed to the more plausible claim that positive conceivability entails negative conceivability, [52] so (CP+) stands in a better chance of being correct than (CP-).

2.5. The Two-Dimensional Framework

In this section I would like to extend and clarify the main tenets of the framework that was introduced in the previous section. The main idea, namely, that of a sentence being evaluated at a possible world considered as actual, was already introduced there. But here I would like to offer a more systematic picture of this framework and introduce some terminology that will be useful later.

Let's recall our examples of a posteriori necessities:

[50] See footnote 41.

[51] For (CP-) says that negative conceivability entails 1-possibility, and we are assuming that 1-possibility entails positive conceivability, therefore it follows that negative conceivability entails positive conceivability.

[52] For (CP+) says that positive conceivability entails 1-possibility, and we are assuming that 1-possibility entails negative conceivability, therefore it follows that positive conceivability entails negative conceivability. As we explained in footnote 41, this seems like a plausible claim.

(1) Water is H_2O

(2) Heat is molecular motion

As we have seen, sentences (1) and (2) are necessary in some sense but not in another: they are true in every possible world considered as counterfactual, but not in every possible world considered as actual. We can say that they are *secondarily necessary*, or 2-necessary, but not *primarily necessary*, or 1-necessary. In other words, their negations, (~1) and (~2), are 2-impossible, that is, they are not true at any possible world considered as counterfactual, but they are 1-possible, as we have seen above, since they are true at some possible worlds considered as actual.

We have seen that (primary) conceivability cannot be a guide to secondary possibility, but it might be a guide to primary possibility, for all we have said so far. I will say a bit more about what the difference between primary and secondary possibility amounts to.

This distinction presupposes that competent speakers are able to consider possible worlds in two different ways. We can consider them as counterfactual possibilities, or as ways the actual world might be. We can also evaluate the extensions that our terms might have in different possible worlds considered as actual and as counterfactual. We seem to have intuitions concerning what the extensions of our terms would be, in those different scenarios. We can say that these intuitions provide us with a way of matching terms with extensions in different possible worlds, considered both as actual and as counterfactual. We can characterize these "matchings" a bit more precisely, in terms of functions or *intensions* that map possible worlds into extensions. For instance, we can say that the term 'water' is associated with an intension that maps each possible world to the extension that the term would have, had that possible world turned out to be the *actual* one. We can call this function the *primary intension* of 'water'. We can also associate the term with a *secondary intension*, which maps every possible world to the extension that the term would have in that possible world considered as *counterfactual*. [53]

[53] This two-dimensional framework has its source in Stalnaker (1978) and Davies and Humberstone (1980), among others. The distinction between primary and secondary intensions is clearly explained in Chalmers (1996) and elsewhere. There are many interesting questions concerning two-dimensional semantics that we cannot discuss here. One such issue concerns how to characterize primary intensions, and their role in explaining notions such as meaning, mental content and a priori knowledge. For

As we have seen above, the primary intension of 'water' would map the actual world to H_2O, and Twin Earth to XYZ, since these are the respective referents that the term would have in those possible worlds considered as actual. We can imagine other possible worlds and see what referent 'water' would have, had these worlds turned out to be actual. (Maybe there are some possible worlds considered as actual where the reference of 'water' is not determinate: this is compatible with this framework. Also, the term could be empty at some scenarios, for instance if watery stuff turned out not to be a single natural kind.) Our intuitions about what 'water' would refer to in all these different possible worlds considered as actual yield the primary intension of 'water': a function from possible worlds considered as actual to extensions.

Likewise, the term is associated also with a secondary intension, that is, a function from possible worlds considered as counterfactual to extensions. This intension cannot be known a priori, since we have to know what the actual world is like, in order to know what 'water' would refer to in different possible worlds considered as counterfactual. Once we know that 'water' refers to H_2O in the actual world, we can see that 'water' will refer to the same entity in every possible world considered as counterfactual, since such a term is a rigid designator. Therefore, the secondary intension of 'water' maps every possible world to H_2O.

We can also evaluate the primary and secondary intensions of whole sentences. In this case, the intensions will be functions from possible worlds to truth-values. The primary intension will be a function from possible worlds to the truth-value that the sentence would have in those possible worlds considered as actual, and the secondary intension will be a function from possible worlds considered as counterfactual to the corresponding truth-values. For example, let's consider sentence (1) again:

(1) Water is H_2O

The primary intension (1-intension) of (1) is a function from possible worlds to the truth-values that (1) would have in each of them, considered as actual. For instance, (1) is true at the actual world, so the 1-intension of (1) would map the actual world to

---

detailed discussion of some of those issues see Chalmers (2004) and the papers in García-Carpintero and Macià (2006). Here, we will mainly focus on the elements of the framework that are directly relevant for the conceivability argument.

the truth-value TRUE, or T for short. But (1) is false at Twin Earth considered as actual, so the 1-intension would map Twin Earth to FALSE, or F for short. (We can also say that the 1-intension of (1) is true at the actual world, and false at Twin Earth.)

We can now characterise primary possibility and necessity a bit more precisely: we can say that a sentence S is primarily possible (1-possible) if and only if its primary intension is true at some possible world (i.e. if and only if S is true at some possible world considered as actual). And S is 1-necessary if and only if its primary intension is true in every possible world (i.e. if and only if S is true in every possible world considered as actual).

Likewise, we can say that S is 2-possible if and only if S's secondary intension is true at some possible world (i.e. if and only if S is true at some possible world considered as counterfactual); and S is 2-necessary if and only if S's secondary intension is true in every possible world (i.e. if and only if S is true in every possible world considered as counterfactual).

The so-called *two-dimensional framework* has it that every term is associated with two intensions: a primary and a secondary intension. Our intuitions as competent users of the terms yield the different extensions in different possible worlds, and therefore we can formulate the corresponding primary intensions.

Chalmers and Jackson employ this framework to analyse Kripke's standard cases of a posteriori necessities.[54] They explain a posteriori necessity in this way: A sentence is a priori when its 1-intension is necessary, that is, when it is true in every possible world considered as actual. And the sentence will be a posteriori when the 1-intension is not necessary, that is, when it is false at some possible world considered as actual. A sentence will be necessary (in the standard sense of 'necessity') when it is 2-necessary, that is, when it is true in every possible world considered as counterfactual.

We can apply this model to sentence (1). This sentence is a posteriori true, and therefore, according to the two-dimensional model, it is not 1-necessary, that is, it is false at some possible world consider as actual, for instance, Twin Earth. But it is a necessary truth (in the standard sense), that is, it is true in every possible world considered as counterfactual, including Twin Earth.[55]

---

[54] See specially Chalmers (1996) and Jackson (1998a).

[55] We can say a bit more about why (1) would be true at Twin Earth, considered as counterfactual. This has to do with our intuitions concerning what entity is relevant for evaluating the truth-value of (1) at

As we can see, Chalmers and Jackson explain a posteriori necessities in terms of sentences that are associated with two intensions: one is necessary (the secondary intension), the other is not necessary (the primary intension).

From this framework, it follows that if a sentence is a posteriori, then its primary intension is not necessary. That is, if the primary intension is necessary, then the sentence has to be a priori. We can express this crucial idea as follows:

(2D) A sentence S is a priori if and only if S's 1-intension is necessary.

This is the crucial claim of the two-dimensional framework. As we can see, this claim explains the aposteriority of a sentence in terms of such sentence having a primary intension that is false at some possible world. This claim has been crucial for the development of conceivability arguments. We can see how (2D) *entails* the conceivability-possibility links that we have discussed above. Let's start with the inference from negative conceivability to possibility.

(CP-) Ideal negative conceivability entails primary possibility.

I will now explain how (CP-) follows from (2D).

Recall that a sentence S is (ideally) negatively conceivable just in case S is not a priori false, that is, just in case ~S is not a priori true. Therefore, if S is (ideally negatively) *conceivable*, then ~S is not a priori true. But if ~S is not a priori true, then according to (2D), its 1-intension will not be necessary. That is, ~S will be false at some possible world considered as actual, that is, S will be true at some possible world considered as actual. That is to say, when S is (ideally negatively) conceivable, S will be true at some possible world considered as actual. This is what (CP-) claims.

Twin Earth considered as counterfactual. Our intuitions say that when we describe Twin Earth as a counterfactual possibility, we would still use the term 'water' to refer to H_2O, not XYZ: we would say, for instance, that in Twin Earth, water is black and tarry, whereas the watery stuff there is not really water but XYZ. We would talk about what happened in that scenario, in this way: *water* turned out to be black and tarry, whereas the stuff that looks like water is not *water*, but XYZ. In this description, 'water' is obviously used to refer to H_2O. This is, according to Kripke (and most contemporary philosophers of language) the way in which we use natural kind terms when describing counterfactual possibilities; this is why these terms are *rigid designators*. Therefore, when we evaluate the truth-value of (1) at Twin Earth considered as counterfactual, the referent of 'water' is H_2O, and therefore (1) comes out as true.

Therefore, the crucial claim of the two-dimensional framework gives support to our two conceivability-possibility links. In particular, it entails that (CP-) is correct, and if (CP-) is correct, then the claim (CP+) is correct too:

(CP+) Ideal positive conceivability entails primary possibility.

This is so because if a sentence is positively conceivable, then it is negatively conceivable. And by (CP-), if it is negatively conceivable, then it is 1-possible. So it follows that if it is positively conceivable, it is 1-possible. That is, (CP+) follows.

Hence, the crucial claim of the two-dimensional framework (2D) has two crucial consequences: (CP+) and (CP-). Therefore, by examining whether these conceivability-possibility links are correct or not, we are examining whether the core claim of two-dimensional semantics is correct or not, since we will be examining whether those important commitments of (2D) are correct or not.

2.6. The Two-Dimensional Argument against Physicalism

In the previous sections we have introduced the notions of conceivability and the conceivability-possibility links that are invoked in the conceivability argument. As we have seen, the argument can be formulated in terms of two different notions of conceivability (ideal *negative* primary conceivability and ideal *positive* conceivability), which in turn give rise to two corresponding conceivability-possibility links, namely, (CP+) and (CP-). I have explained how the two-dimensional framework, and in particular the two-dimensional account of a posteriori necessities, can give support to both conceivability-possibility theses.

Therefore, we can formulate two different versions of the conceivability argument, the first one appealing to (CP+), and the second one appealing to (CP-), as follows:

The Conceivability Argument ((*CP+*) *version*)
CA+1: P&~Q is positively conceivable.[56]
CA+2: If P&~Q is positively conceivable, then it is primarily possible.

[56] In what follows, we will assume that we are talking about ideal primary conceivability.

CA+3: If P&~Q is primarily possible, then physicalism is false.

_____

CA4: Physicalism is false.

The Conceivability Argument ((*CP-*) *version*)

CA-1: P&~Q is negatively conceivable.

CA-2: If P&~Q is negatively conceivable, then it is primarily possible.

CA-3: If P&~Q is primarily possible, then physicalism is false.

_____

CA4: Physicalism is false.

We can see that the second premise of the first argument, that is, CA+2, is just an instance of the thesis (CP+), and that the second premise of the second argument, that is, CA-2, is just an instance of the thesis (CP-). As we have seen, theses (CP+) and (CP-) are consequences of the two-dimensional framework, and in particular of the core thesis of two-dimensional semantics, namely, (2D) above. These two theses postulate two different conceivability-possibility links, which, as Chalmers and Jackson have argued, can overcome the standard Kripkean counterexamples to the original conceivability-possibility link. Both theses are compatible with there being sentences that are conceivable but not secondarily possible, to the extent that these sentences are primarily possible. Therefore, we have arrived at two considerably stronger versions of the conceivability argument, because they are not affected by one of the main obstacles for any conceivability-possibility inference, namely, the aforementioned Kripkean cases of a posteriori necessities.

Since these versions of the conceivability argument appeal to two-dimensional semantics in order to solve the problems posed by the familiar Kripkean counterexamples, they are also known as (versions of) *the two-dimensional argument* against physicalism. As we have seen, the second premise of each argument makes use of the crucial idea of two-dimensional semantics, namely, that a conceivable sentence has to be true in at least some possible world considered as actual (although not necessarily considered as counterfactual).

In this section, I will lay out the two-dimensional argument in considerable detail, paying special attention to some problems that might arise, in order to arrive at

the strongest version of the argument that is available. I will particularly focus on Chalmers' presentation and discussion of the argument, since it the most elaborated.[57]

Before turning to that, some clarificatory comments are in order. I have emphasized above that (CP-) is a more ambitious thesis than (CP+), since the former has certain controversial commitments that the latter does not have (namely, the former is committed to negative conceivability entailing positive conceivability). Therefore, someone might argue that we should get rid of (CP-), and focus instead on (CP+), since there are some putative counterexamples that might affect the former which will not affect the latter. This is true: the premise CA+2 has a better chance of being true than its counterpart CA-2. However, on the other hand, the corresponding premise CA-1 seems more plausible than its counterpart CA+1. In particular, the negative conceivability of P&~Q is easier to establish than its positive conceivability. For it is always easier to show that a certain sentence cannot be ruled out a priori, than showing that we can imagine a perfectly coherent scenario where such sentence is true. Because of this, both versions of the conceivability argument deserve attention. The version based upon positive conceivability has the virtue that its conceivability-possibility link CA+2 is (perhaps) more reliable than its counterpart, although it has the slight disadvantage that its conceivability claim CA+1 is harder to establish. On the other hand, the version of the conceivability argument based upon negative conceivability puts forward a (perhaps) slightly bolder conceivability-possibility link CA-2, but it has the advantage that its conceivability claim CA-1 is easier to establish. Therefore, both formulations have their vices and virtues.

However, many of the things we will have to say in what follows apply to both versions of the conceivability argument, so sometimes I will speak just of 'conceivability', as a neutral term for both positive and negative conceivability (though in most cases, I will be referring to ideal primary conceivability). When the distinction between negative and positive conceivability becomes relevant, I will make it explicit. Otherwise, we will just talk about conceivability in general, bearing in mind that what we say can be applied to both notions of conceivability.

Let us turn now to the third premise of the two conceivability arguments above. That premise claims that if P&~Q is primarily possible, then physicalism is

[57] The first detailed presentation of this argument can be found in Chalmers (1996). The two-dimensional argument is further developed in his (2002a), (2003) and especially in his (forthcoming), from which I draw most of what follows.

false. This premise does not seem very plausible, for the following reason. Let's recall the thesis of Physicalism, as we characterized it in the first chapter:

> **Physicalism** is true if and only if any possible world *w* which is a *minimal* physical duplicate of @ is a duplicate *simpliciter*.

It can be argued at this point that the primary possibility of P&~Q does not have to pose a problem for materialism, because, the objection goes, only secondary possibility is genuine metaphysical possibility. Physicalism, and in particular the claim of global supervenience that we have used to capture the intuitive idea of physicalism, is concerned with what is the case in possible worlds considered as *counterfactual*, that is, the kind of possibility that we have called 'secondary possibility'. What is at issue here is whether all *counterfactual* possible worlds where P is the case are worlds where Q is the case too. Therefore, to establish the primary possibility of P&~Q is not sufficient, *per se*, for establishing its metaphysical possibility, since the latter involves the secondary possibility of P&~Q.

Chalmers' response to this objection is twofold. On the one hand, it can be argued that the primary possibility of P&~Q entails its secondary possibility. In order to establish that, we should show that the evaluation of P (the physical description of the world) and Q (any phenomenal truth) in worlds considered as actual corresponds to the evaluation of P and Q in possible worlds considered as counterfactual. That is, we should show that P's 1-intension and 2-intension coincide, and likewise for Q's 1-intension and 2-intension. Is this plausible? This depends on whether the terms involved in P and Q, that is, physical terms and phenomenal terms, respectively, have the same referents in worlds considered as actual and as counterfactual.

There are terms, though, that are evaluated differently in a world considered as actual and considered as counterfactual, for example, 'water': it refers to XYZ in Twin Earth considered as actual, whereas it refers to H_2O in Twin Earth considered as counterfactual. That is to say, the primary and secondary intensions of 'water' do not coincide. The reason for this divergence is that it makes sense to distinguish between the way in which water looks and water itself: we can imagine stuff that look like water but is not water (XYZ), and water that does not look like water (for example, H_2O in Twin Earth, where it is black and tarry). When we evaluate the reference of 'water' in a world considered as actual, we search for the stuff that *looks like* water

there (the watery stuff), whereas when we evaluate the reference of 'water' in a world considered as counterfactual, we look for *actual* water (H_2O); and these two things can differ in a given world.

But it seems that there is no similar distinction for the case of phenomenal terms, such as 'pain'. Arguably, we cannot distinguish between what feels like pain and what is pain, since for something to be pain is just for it to feel like pain. There is no distinction between appearance and reality for phenomenal properties. We cannot imagine things that look like pain but are not real pain, nor things that are pain but do not feel like pain.[58] As a consequence, 'pain' has the same reference in worlds considered as actual and as counterfactual, that is, its primary and secondary intension coincide. Analogous reasoning can show that the same applies for all phenomenal terms. Therefore, sentences involving phenomenal terms, such as Q, will be evaluated analogously in worlds considered as actual and as counterfactual.[59]

However, *pace* Kripke, someone could have a view about the reference of phenomenal terms on which their primary and the secondary evaluations differ. For instance, someone could argue that, say, 'pain' refers, in possible worlds considered as actual, to the property of feeling like pain, and, in counterfactual worlds, to the property that actually realizes pain in human beings. Anyway, Chalmers argues,[60] even if we accept this view of the reference of phenomenal terms, we could run a similar conceivability argument using a new description of phenomenal facts: not Q (for which, according to this view, primary and secondary intensions differ), but Q*, for which, we can stipulate, the 1-intension and the 2-intension are the same, namely, Q's 1-intension. For instance, if Q is a truth about pain, Q* would be a truth about painful states (states that feel like pain, regardless of what realizes them). Therefore, if Q* is true at some possible world considered as actual, it will be true at that world considered as counterfactual, since its primary and secondary intensions are the same.

To sum up: we have seen that there are good reasons to think that for Q, the primary and secondary intensions coincide. And it is also plausible that for P, the complete physical description of the world, that is the case too. (Although we will explore later what would follow if that was not the case for P.) With these two assumptions in place, we can run a new version of the conceivability argument, as

[58] This has been argued by Kripke (1980: 331).
[59] This point is also made in Chalmers (1996: 133) and elsewhere.
[60] See Chalmers (1996: 133-4).

follows. (We will call this the two-dimensional argument against physicalism, since it is based on the two-dimensional framework discussed above.)

The Two-Dimensional Argument

2DA1: P&~Q is conceivable.

2DA2: If P&~Q is conceivable, P&~Q is primarily possible.

2DA3: If P&~Q is primarily possible, P&~Q is secondarily possible.

2DA4: If P&~Q is secondarily possible, Physicalism is false.

_____

2DA5: Physicalism is false.

We can see that the fourth premise, that is, 2DA4, clearly follows from our characterization of Physicalism above, since this is committed to there being no minimal physical duplicate of the actual world that is not a duplicate simpliciter. But if P&~Q is secondarily possible, then there is a minimal physical duplicate that is not a duplicate in all respects. So Physicalism would be false.

The motivation for the third premise, that is, 2DA3, is that it is plausible to assume that the primary and secondary intensions of P coincide, and likewise for the primary and secondary intensions of Q. Both assumptions could be denied, but in each case a new version of the two-dimensional argument can be formulated, as we will see, so those assumptions are not really essential for the success of the conceivability argument. Let's examine in turn what these two different versions of the argument would be like.

We have already seen how to respond to the worry that Q's primary and secondary intensions could not coincide, after all. We could just introduce a new sentence, Q*, which by definition would have Q's 1-intension as both primary and secondary intensions. Then, the two-dimensional argument would run as follows:

The Two-Dimensional Argument (Q Version)*

2DA1: P&~Q is conceivable.

2DA2: If P&~Q is conceivable, P&~Q is primarily possible.

2DA3*: If P&~Q is primarily possible, P&~Q* is secondarily possible.

2DA4*: If P&~Q* is secondarily possible, Physicalism is false.

_____

In this way, we can see that the primary possibility of P&~Q is sufficient to pose trouble for physicalism. Remember that we are assuming for the moment that P's primary and secondary intensions do coincide. (We will explore later what would follow if we deny this assumption.) If that is the case, then even if the primary possibility of P&~Q does not imply its secondary possibility, it does imply the primary, and therefore, by definition, the secondary, possibility of P&~Q*. And this is enough to put physicalism in jeopardy, since it would imply that certain facts, those referred to in Q*, do not globally supervene on the physical, since there is a possible counterfactual world where P holds but Q* does not.

Therefore, the assumption that Q's primary and secondary intensions are the same is not really essential for the two-dimensional argument to go through. In any case, it is a plausible assumption, as we have said above. For the sake of exposition, we will typically focus on the standard version of the two-dimensional argument, but we can see that there is another version of the argument that does not need that assumption, which could be deployed if the claim that Q's primary and secondary intensions coincide proved to be an unwarranted assumption.

What about the assumption that the primary and secondary intensions of P are also the same? Is this really plausible? Remember that P is a complete physical description of the world.

There is a view about the semantics of physical terms according to which the primary and secondary intensions of such physical terms do not coincide. This view could pose a problem for the third premise above, so it would be useful to discuss it here.

According to this view, the information that physical theory provides about physical entities is mainly information about their *relations* with other physical entities. In this way, physical theory offers characterizations of physical entities in terms of the different *roles* they play: we can label these the *theoretical roles* of physical entities. A crucial question is: how do these physical terms pick up the referents they have? According to the view under discussion, they fix their referents by means of those theoretical roles. In particular, this view has it, the corresponding *primary intensions* of those physical terms are related to the theoretical role that the corresponding physical entities are supposed to play in physical theory.

A crucial idea of this view is that those theoretical roles are realized by, or grounded in, certain *intrinsic properties* of the corresponding physical entities. And, this view claims, the corresponding physical terms refer to the intrinsic properties that realize those theoretical roles. That is, physical terms refer to intrinsic properties, *via* the theoretical roles that are associated with each physical term. [61]

More precisely: For a given physical term G, the primary intension will refer to whatever satisfies G's theoretical role, but the secondary intension will refer to the entity or property that *actually* satisfies such theoretical role. That is, G will refer, in different possible worlds considered as actual, to the different properties that satisfy G's theoretical role, but G will refer, in different possible worlds considered as counterfactual, to the very same property that satisfies G's theoretical role at the world that we have fixed as actual. A crucial assumption of this view is that such theoretical roles can be satisfied by *different* intrinsic properties in different possible worlds. Therefore, it follows that G's referent in a given possible world considered as actual could differ from G's referent in the same possible world consider as counterfactual. That is, G's primary intension and secondary intension do *not* coincide.

For example, let's imagine that the intrinsic property that satisfies G's role is X in the actual world, and Y in some other counterfactual possible world W. Then, G's 1-intension refers to Y at W, since Y satisfies the corresponding theoretical role at W, but the 2-intension refers to X, since X satisfies that theoretical role at the actual world. Therefore, on this view, the primary intension and secondary intension of a physical term do not coincide. So this view could bring problems for the third premise above, namely, 2DA3.

What 2DA3 says is that if P&~Q is true at some possible world considered as actual, then it is true at some possible world considered as counterfactual. What would be required for this premise to fail?

Firstly, we would need to endorse the view about the semantics of P that we just explained. According to this view, P could be true at a possible world W considered as actual, but false at W considered as counterfactual. In this way, this possible world W would instantiate the same theoretical roles as the actual world, but with different intrinsic properties playing the corresponding theoretical roles.

[61] See Chalmers (1996: 153-5) and (2003: 117) for further discussion of this view.

What would take for a world to satisfy P&~Q considered as actual, but not when it is considered as counterfactual? That is, what would take for a world to satisfy the 1-intension of P&~Q but not the 2-intension, according to this view about the semantics of P? This would mean that the 1-intension of P holds but the 1-intension of Q does not hold. That is, in that case the *relational* character of physical entities (the theoretical roles of such physical entities) does not determine Q (the phenomenal properties that are instantiated in a world). On the other hand, the fact that the 2-intension of P&~Q does not hold in any world means that the 2-intension of P does determine Q. That is, what determines whether Q is instantiated or not is the *intrinsic* character of the physical world.

Therefore, according to this view there is a possible world W where the 1-intension of P holds but the 2-intension does not hold, and therefore the phenomenal truths might vary. That is, Q might be false, since the presence of different intrinsic properties might give rise to different phenomenal properties. Therefore, W can be a world where P&~Q is true when considered as actual, but false when we consider it as counterfactual, since the 2-intension of P does not hold at W.

This view is labelled *Russellian monism* (or *Type-F monism*), and, as Chalmers argues, it is the view that would make the third premise above false. According to this view, there are possible worlds considered as *actual* where P is true but Q is false. But there are no worlds considered as *counterfactual* where P is true and Q is false, because these worlds have the same intrinsic properties as the actual world and therefore the same phenomenal properties.

Therefore, for Russellian (or type-F) monism, the fact the P&~Q is true at W considered as *actual* does not entail that P&~Q is true at W considered as *counterfactual*. So, this position would entail that the third premise of the two-dimensional argument above, that is, 2DA3 (and also 2DA3*) is false. The crucial question now is: is this view a physicalist theory?

According to Russellian monism, the relational properties that physical theory tells us about, that is, the primary intensions of physical terms, do not determine what phenomenal properties are instantiated in a world. There are worlds where these relational properties are the same and still, Q is false. What we need in order to fix the phenomenal character of a world is to fix the intrinsic properties that realise the theoretical roles that physics talks about. Chalmers suggests that Russellian Monism is more similar to dualist theories than to physicalist ones. The idea is that what

physics tells us about the world (the relational properties that are instantiated) is not sufficient to fix the phenomenal character of the world: we need to invoke intrinsic properties of physical objects, which are responsible for the instantiation of phenomenal properties. This theory seems to have something in common with anti-physicalist views.

In any case, this issue is not completely clear to me. After all, we have seen that Russellian monism asserts that physical terms do refer to intrinsic properties, by virtue of their satisfying certain theoretical roles. So, in a sense, P does refer to those intrinsic properties that are responsible for the instantiation of Q. Indeed, all possible worlds where the secondary intension of P holds are worlds where Q holds too. So perhaps Russellian monism can be seen as a physicalist view after all.[62]

By way of conclusion, we can say that Russellian monism poses a challenge to the inference from the 1-possibility of P&~Q to its 2-possibility. Indeed, Russellian monism looks like the only view which can have that consequence. Hence, Chalmers puts forward a new formulation of the argument, to overcome this problem:

The Two-Dimensional Argument (Final Version)

(1): P&~Q is conceivable.

(2): If P&~Q is conceivable, P&~Q is primarily possible.

(3): If P&~Q is primarily possible, either P&~Q is secondarily possible or Russellian (type-F) monism is true.

(4): If P&~Q is secondarily possible, Physicalism is false.

———————

(5) Either Physicalism is false or Russellian (type-F) monism is true.

In this final version of the two-dimensional argument, the possibility of Russellian monism being true is explicitly acknowledged. As we have seen, unless this view is true, it will follow that if P&~Q is 1-possible, it will be 2-possible as well. So (5) is the proper conclusion of the two-dimensional argument: either P&~Q is 2-possible and therefore physicalism is false, or Russellian monism is true. For Chalmers, in

[62] I think that this is a very interesting issue, and more research on it could prove to be very fruitful. Daniel Stoljar has prominently developed this account and has argued that it can be seen as a defence of physicalism. (See Stoljar (2001a), (2001b) and (2006).) Whether this sort of view is ultimately successful is something that we cannot determine here, although I do think that this strategy deserves serious attention.

either case physicalism would be in trouble, because he thinks that Russellian monism is not compatible with physicalism, as intuitively conceived. In any case, the final version of the two-dimensional argument is neutral on this question.

So one possible defence of physicalism would be to argue that Russellian monism is true and that it is actually a species of physicalism. This is a very interesting project, but it is not the one that I will pursue in what follows. Rather, my aim is quite different: what I want to argue is that the conceivability of P&~Q does not entail its 1-possibility. Therefore, we can put aside the issue of whether the 1-possibility of P&~Q would really entail its 2-possibility. For I will argue that we can block the conceivability argument at an earlier stage: premise (2) above is incorrect, or so I will argue.

To sum up: in this chapter, we have introduced the main versions of the conceivability argument, and we have examined in great detail the most sophisticated version of it available, namely, the two-dimensional argument, which poses one of the main contemporary challenges to physicalism. In this thesis, we will argue that this argument ultimately fails, because the second premise (the conceivability-possibility link) is wrong. (This response is compatible with other responses to the conceivability argument being correct as well, although we will not examine them here.)

3. Epistemic Arguments against Physicalism II

In this chapter, we will explore some other important epistemic arguments against physicalism that have been widely discussed in contemporary debates. We will start by introducing two important epistemic arguments that can be seen as precursors of the conceivability argument. These are the property dualism argument, defended by Stephen White (1986), and the modal argument, presented in Saul Kripke (1980). Then we will turn to another influential epistemic argument that was also mentioned at the beginning of chapter 2, namely, the knowledge argument. With this we will complete our introduction to epistemic arguments against physicalism. [63]

The main aim of this discussion of these other epistemic arguments is twofold: firstly, I want to show that the standard formulations of these epistemic arguments face important problems. Secondly, we will see that it is possible to capture the core ideas of those arguments in terms of the two-dimensional framework that we explored in the previous chapter. This will lead us to the formulation of a new two-dimensional argument, namely, the two-dimensional argument against *type-identities*. This will set up the agenda for the following chapters, which will be concerned with the criticism of both two-dimensional arguments: the one against Physicalism and the one against type-identities.

We will finish this chapter with an important distinction between two ways of responding to these two-dimensional arguments and epistemic arguments against physicalism in general. The first response has it that the epistemic premise of each epistemic argument is false (that is, the alleged epistemic gap between the physical and the phenomenal does not really hold). The second response has it that although that epistemic premise does hold, it does not really entail the metaphysical conclusion it is supposed to entail. We will briefly discuss the first response before turning to the second one, which will be our main focus in this thesis.

---

[63] There is another epistemic argument that is also discussed in the literature, namely, the so called *explanatory gap* argument. In a nutshell, this argument has it that, since physical truths cannot entail phenomenal truths a priori, then we cannot offer an *explanation* of phenomenal truths in physical terms. Most versions of the explanatory gap argument (such as Levine (1983), (1993)) differ from the other epistemic arguments that we are examining here in that, whereas the explanatory argument has an *epistemic* conclusion (i.e. consciousness cannot be *explained* in physical terms), the other arguments have *metaphysical* conclusions (e.g. consciousness is not physical). Here, we will be mainly concerned with the second sort of arguments, since we are mainly interested in the question of whether we can draw metaphysical conclusions from epistemic premises. Block & Stalnaker (1999) offer a very detailed discussion and criticism of the explanatory gap argument.

3.1. The Property Dualism Argument

The property dualism argument was first discussed by J.J.C Smart (1959)[64], and has been further developed by Stephen White[65], among others.

White starts his presentation of the argument by assuming token-identity between physical and phenomenal states, such as the following:

(A): Smith's brain state X at *t* is Smith's pain at *t*

As we noticed in Ch. 1, the view that every mental token is identical with some physical token has played an important role in the discussion of physicalism. It is an application of the view that we labelled 'token-physicalism', according to which every token that is instantiated in the actual world is identical to some physical token. Token-physicalism is usually taken to be weaker than the thesis of type-physicalism, according to which every type, and therefore every mental type (say, *being pain*), is identical to a physical type (say, *being C-fibre firing*).[66] However, as we explained in Ch. 1, Physicalism, as we have characterised it here, is not committed to the truth of token-identity: it is merely committed to the global supervenience of all facts on physical facts[67], but this does not have to commit one to holding that every mental token is identical to some particular physical token. In any case, token-identity is widely held among physicalists, and it will be interesting to see what follows from that.

White then asserts that the truth of the identity statement (A) cannot be established a priori: "the fact that ['Smith's brain state X at *t*' and 'Smith's pain at *t*'] are coreferential cannot be established on a priori grounds" (1986: 705). This seems obvious. From this, he infers the following: "Thus, there must be different properties of Smith's pain (i.e. Smith's brain state X) in virtue of which it is referent of both terms" (705-6).

---

[64] Smart discusses this argument under 'Objection 3' (see Smart (1959: 63-4)), and gives the credit for this objection to Max Black. He also says that "it is the most subtle of any of those I have considered, and the one which I am least confident of having satisfactorily met" (1959: 68, fn. 11).
[65] See White (1986) and (forthcoming).
[66] A classical source is Davidson (1970); see also Fodor (1974).
[67] See Ch. 1 for a more precise characterization of Physicalism.

This inference relies on an account of a posteriori identities known as the Distinct Property Model (DPM)[68]. According to this model, any given a posteriori identity statement 'A=B' is a posteriori because the terms 'A' and 'B' pick out the referent by virtue of the different properties associated with each term. We can put the main idea as follows:

> (DPM): For every a posteriori identity statement 'A=B', each term refers to the referent by virtue of a distinct property associated with the term.

This model relies on a theory of reference for singular and general terms which we can call the "mode of presentation" theory of reference, and which claims the following:

> *Mode of Presentation Theory*: Singular and general terms are a priori associated with modes of presentations that determine their referents: these modes of presentation consist of a *description* that is a priori coreferential with the term, and a corresponding *property* by virtue of which the term picks out its referent.

According to White, every mode of presentation (MOP) has two sides: the linguistic side, that is, a description associated with the singular term that is a priori coreferential with it, and the object's side, that is, a property of the referent by virtue of which the singular term refers to that referent. The former is called the representational MOP, while the latter is called the non-representational MOP.

One of the main motivations for this theory is that it seems to fit most standard cases. Let's explain this with an example. The name 'Phosphorus' refers to Venus. 'Hesperus' also refers to Venus. The following identity statement is a posteriori:

> (B) Hesperus is Phosphorus

White argues that, since the two names are a posteriori coreferential, they have different descriptions associated a priori with them. We can say that the descriptions

[68] This label is introduced in Levine (2001).

associated a priori with 'Hesperus' and 'Phosphorus' are, respectively: 'the first heavenly body visible in the evening' and 'the last heavenly body visible in the morning'. These two descriptions are not a priori linked.

White explains that these two descriptions *express* two different properties, by virtue of which each term refers to its referent. That is, each term expresses a different reference-fixing property: 'Hesperus' refers to Venus because Venus exemplifies the property *being the first heavenly body visible in the evening*, and 'Phosphorus' refers to Venus because Venus exemplifies the property *being the last heavenly body visible in the morning*.

In general, this theory holds that when we have two singular terms e_1 and e_2 that are a posteriori coreferential, they refer to the same object O in virtue of two different modes of presentation, that is, they have associated two different descriptions D_1 and D_2 that correspond to two different properties of the referent P_1 and P_2 by virtue of which the two terms refer to O. (We can say that these two properties are expressed by e_1 and e_2 respectively).

We can also apply this model to identity statements involving general terms, such as the following:

(1) Water is H_2O
(2) Heat is molecular motion

In the first case, the terms 'water' and 'H_2O' refer to the same referent, that is, H_2O, by virtue of two distinct properties: 'being watery stuff', the property associated with the term 'water', and the property 'having the molecular composition of H_2O', the one associated with 'H_2O'. In the second case, the properties are 'being the phenomenon that produces sensations of heat' (associated with 'heat') and 'being molecular motion' (associated with 'molecular motion').

How does the Distinct Property Model work in those cases? How can we explain the aposteriority of those identity statements in terms of two distinct properties of the referent? As Joseph Levine puts it, "in […] these cases, it's easy to see how our empirical discovery that they are the same thing (or stuff) is a matter of discovering that this one thing (or stuff) has different properties that we originally thought might be instantiated by different things" (Levine (2001:47)).

Therefore, if we accept that the identity statement (A) is a posteriori, and endorse the account of a posteriori identities (DPM), then it follows that the expressions 'Smith's brain state X at *t*' and 'Smith's pain at *t*' refer to their referent by means of different properties. What follows from this? According to White, "since there is no physicalist description that one could plausibly suppose is coreferential a priori with an expression like 'Smith's pain at *t*', no physical property of a pain (i.e. a brain state of type X) could provide the route by which it was picked out by such an expression" (1986: 706). That is, White is assuming that there is no physical expression that is a priori corefential with the expression 'Smith's pain at *t*'. Therefore, he argues, the property by virtue of which 'Smith's pain at *t*' picks out its referent must be different from any physical property. For if it was identical to any physical property, then there should be a physicalist expression that was coreferential with 'Smith's pain at *t*', that is, there should be a corresponding psychophysical identity statement that was a priori. Since there is no such thing, then 'Smith's pain at *t*' does not refer to its referent by virtue of a physical property.

Is there another option? White also discusses the possibility of a topic neutral property being the reference-fixing property: "Let us stipulate that a property which is neither physical nor mental is topic neutral" (706). Functional properties, for instance, can be seen as topic neutral. So the proposal here is that 'Smith's pain at *t*' is a priori coreferential with some topic neutral description, such as "an event like the event that occurs in one when one is stabbed with a pin (and so forth)" (706) (this proposal was suggested by Smart (1959) as his solution to the property dualism argument). But this idea is controversial for familiar reasons: we can always conceive of a being that satisfies the topic neutral description but not the phenomenal one, and therefore, it would follow that phenomenal terms are not a priori coreferential with such topic neutral terms.[69] So, what follows from this? White argues as follows:

> If there are no topic neutral expressions which are at least coreferential a priori with such mentalistic descriptions as 'Smith's pain at *t*', then these mentalistic descriptions refer in virtue of a property distinct from that in virtue of which any physicalistic or topic neutral

[69] Interestingly, White (1986) put forward the Property Dualism Argument as an argument for a functionalist view: he argued that if there were no physical-functional descriptions associated a priori with terms such as 'Smith's pain at t', then Property Dualism would inevitably follow, and therefore the only solution was to offer some sort of functional analysis of phenomenal terms. However, in his more recent papers, such as (forthcoming) and elsewhere, he claims that such functional analyses are wrong and therefore Property Dualism follows.

> expression refers. Such a property or feature could only be regarded as an irreducibly mental entity. (1986: 707)

The property dualism argument, then, has it that since phenomenal terms such as 'Smith's pain at *t*' are not a priori coreferential with any physical or topic neutral expression, then they pick out their referents by virtue of non-physical, irreducibly mental properties. Therefore, the conclusion is a form of property dualism, that is, the view that phenomenal states have two kinds of properties: physical ones and non-physical, irreducibly mental ones.

We can now put the argument as follows:

The Property Dualism Argument

PDA1: The singular terms 'Smith's brain state X at *t*' and 'Smith's pain at *t*' are not a priori coreferential.

PDA2: If they are not a priori coreferential, then they refer to their referent by virtue of different properties of the referent (by DPM).

PDA3: The description a priori associated with 'Smith's pain at *t*' is neither physical nor topic neutral.

PDA4: If the description a priori associated with 'Smith's pain at *t*' is neither physical nor topic neutral, then the property by virtue of which 'Smith's pain at *t*' refers to its referent is neither physical nor topic neutral.

———————

(PDA5): Smith's pain state at t has both physical properties and non-physical properties. (Property Dualism)

Crucial to this argument are premise PDA2 and the account of a posteriori identities that it relies upon, namely, the Distinct Property Model. As we will see later, this model is closely related to the two-dimensional framework, and they both share a similar motivation. Therefore, we will examine this step in the argument in some detail. However, before doing that, I would like to discuss some other more specific problems of the property dualism argument. So for the time being, and for the sake of

discussion, we will assume that the DPM is correct and we will examine the remaining premises of the argument.[70]

The reader may be wondering which notion of 'physical' is appealed to in this argument. This is a fair question: unfortunately some advocates of the argument such as White (1986) are not explicit about the answer. Therefore, I will explore the different versions of the argument that result when we use different notions of the physical, in order to see which version is most appropriate.

In Ch. 1, we distinguished between 'physical' in a narrow sense ($physical_n$) and 'physical' in a broad sense ($physical_b$), where the former concerns physical properties and entities of the sort posited in current physical theory, while the latter concerns those properties globally supervening on the former. Therefore, if we want the conclusion of the property dualism argument to pose any problem for physicalism, then the notion of 'physical' involved must be '$physical_b$'. It would not be a problem for physicalism to say that pain states have properties that are not posited by physical theory: this is rather trivial. What is at issue is whether pain states have properties that do not globally supervene on the $physical_n$ (in the appropriate sense).

It is clear that if we want the property dualism argument to be valid and not to commit equivocation, the all the uses of 'physical' in the argument should be the same. But as we have seen, any interesting conclusion has to involve the broad notion of 'physical'. What happens if we understand the premises of the argument as involving this sense too? One crucial premise here is PDA3: 'The description a priori associated with 'Smith's pain at *t*' is neither physical nor topic neutral'. How should we understand 'physical' here?

Well, first of all, in this case we are talking about *descriptions*, not properties. What does it mean to say that the *description* associated with a term is physical (or not)? To get clear about this, first we need to get clearer about the MOP theory. In particular, we should be clearer about which descriptions are at issue here. As we have seen before, according to the *Mode of Presentation Theory* our terms are associated with a mode of presentation, which consists of a description a priori associated with the term, and a property of the referent by virtue of which the term picks out the referent. What is the relation between this description, that is, the representational mode of presentation, and the corresponding property, that is, the

[70] We can also assume that premise PDA1 is correct too.

non-representational mode of presentation? White answers this crucial question as follows:

> The representational and the non-representational modes of presentation must be appropriately connected: the connection between the two must be a priori. […] The route to the referent must explain how the subject's term succeeds in picking out the referent in question. The expression 'the morning star', for example, does not pick out Venus in virtue of the property of being the only planet with a maximum surface temperature of 480 ºC. (White (forthcoming: 2))

White holds that that there must be an accessible correspondence between an associated description and the corresponding property. Effectively, the property of Venus by virtue of which 'Phosphorus' refers to Venus is *being the last heavenly body that is visible in the morning*, at least for those subjects that use that type of description to fix the reference of 'Phosphorus'. For most subjects, the reference-fixing property cannot be *being the only planet with a maximum surface temperature of 480º*, because those subjects do not associate a priori the description 'the only planet with a maximum surface temperature of 480º' with the term 'Phosphorus'.

We can encapsulate White's claim about the link between the representational and the non-representational modes of presentation (MOPs) as follows:

> (LINK): For every representational mode of presentation (description 'the x such that F'), there is an associated predicate, 'being the x such that F', that refers to the property of *being the x such that F*. This is the relation between description and property, which must be a priori accessible to the subject.

Following this idea, we can formulate the following schema:

> Singular term: e
> Description D (representational MOP): 'the x such that F'
> Associated Predicate Y: 'being the x such that F'
> Property Z (non-representational MOP): *being the x such that F*

With this schema in mind, we can come back to our initial question, namely, what are *physical descriptions*? The answer to this question depends in part on the

sort of property that is expressed by such a description. (Recall that a description D *expresses* a certain property Z by means of being associated with a predicate Y which *refers* to such a property Z).

In order to answer our question about what physical descriptions are, we can distinguish between physical descriptions in a *narrow* sense ($physical_n$) and physical descriptions in a *broad* sense ($physical_b$). *$Physical_n$ descriptions* would be those descriptions that are invoked in physical theory, whereas $physical_b$ descriptions would be those descriptions *expressing* properties that globally supervene on $physical_n$ properties (that is, $physical_b$ descriptions are those whose associated predicates refer to $physical_b$ properties).

Then, the question at issue now is: Could we understand PDA3 as invoking $physical_b$ descriptions? I think it is clear we cannot, because we cannot just assume that the description that we associate a priori with 'Smith's pain at t' does not express a $physical_b$ property. This would clearly beg the question against the physicalist. Arguably, the only predicates that seem to be a priori associated with 'Smith's pain at *t*' are phenomenal predicates themselves, but what is at issue is precisely whether phenomenal predicates refer to $physical_b$ entities or not. To assume that they refer to non-$physical_b$ properties is to beg the question at issue here. Therefore, the advocate of the property dualism argument cannot just assume that the corresponding descriptions are not $physical_b$, because this would just mean that they express properties that are not $physical_b$, and whether there are any of these properties is precisely what is at stake.

We could avoid this circularity by interpreting PDA3 in terms of $physical_n$ descriptions. In this way, PDA3 would seem justified: it seems plausible to say that the expression 'Smith's pain at *t*' is not a priori coreferential with any description of the sort postulated in current physics (nor with any topic-neutral description). However, if we want PDA3 to be plausible (and PDA5 to be interesting), then PDA3 and PDA5 must involve different notions of 'physical'. So, what happens to PDA4? Can it be plausible, even if it involves different notions of 'physical' in the antecedent and in the consequent? Let's recall PDA4, making explicit the two notions of the physical involved:

PDA4: If the description a priori associated with 'Smith's pain at *t*' is neither $physical_n$ nor topic neutral, then the property by virtue of which 'Smith's pain at *t*' refers to its referent is neither $physical_b$ nor topic neutral.

Is this premise plausible? I think it is not. For the *antecedent* of this conditional has it that the corresponding description does not belong to the set of physical descriptions invoked in current physics, and from this we can just infer that the corresponding property is not $physical_n$, that is, the reference-fixing property for 'Smith's pain at *t*' does not belong to the set of properties posited by current physics. But then, from this we cannot infer that such property is not $physical_b$. We can (perhaps) infer that pain states have properties that are not posited by current physical theory, but this does not imply that these properties do not supervene on $physical_n$ properties. There are many properties that are not posited by current physical theory, for instance the property of being a chair, or the property of being a river. But it seems plausible to say that these properties are $physical_b$.

Therefore, we can conclude that the property dualism argument will not pose serious problems for physicalism, regardless of which notion we use. To sum up the problems: If we invoke the narrow sense ($physical_n$), then PDA3 will be plausible, since the description by virtue of which 'Smith's pain at *t*' refers to Smith's pain is not a description posited by physical sciences. But if we understand PDA5 as just saying that pains have non-$physical_n$ properties, this will not pose a problem for physicalism. On the other hand, if we invoke the broad sense of 'physical', as properties that supervene on $physical_n$ properties, (PDA3) would beg the question. The descriptions associated a priori with 'Smith's pain at *t*' express *phenomenal* properties, but this would imply that such descriptions are not $physical_b$ only if we assumed that phenomenal properties are not $physical_b$. But of course this cannot be assumed in this context. Finally, we could interpret PDA3 according to the narrow sense of 'physical', and PDA5 according to the broad sense. In this way, I think that PDA3 would be true, and PDA5 would entail the rejection of physicalism, but then PDA4 would be incorrect, and therefore the property dualism argument would not work either.

So we have seen that the standard formulation of the property dualism argument is not compelling. However, there is another version of the property dualism argument that might perhaps pose a more significant challenge to physicalist views,

namely, the one based upon the two-dimensional framework. White himself points out this sort of model in a more recent paper.[71] Let us see now how to reconstruct the property dualism argument along these lines. [72]

Let us consider the following type-identity statement:

(C): Being hurtful is being a neurophysiological state of type N

White claims that, when we have an a posteriori identity such as (C), it is always possible for a rational believer to believe the following: "I am in a state that hurts and not in a neurophysiological state of type N" (White (forthcoming: 4)). That is, since (C) is a posteriori, it is possible for a rational subject to believe that it is false, and that she is instantiating one property but not the other. But, according to White, we need an explanation for the fact that such belief is rational, and he suggests that this explanation should involve "a logically possible world at which the subject could be in a state that was hurtful and not be in the neurophysiological state in question" (forthcoming: 3). This sounds very similar to the two-dimensional framework rehearsed in the previous chapter: what White is claiming here is that in order to explain the aposteriority of an identity statement, we need to appeal to a possible world where such statement is indeed false. This seems very close to the core claim of the two-dimensional framework that we explained in the previous chapter. Let's recall it here:

(2D) A sentence S is a priori if and only if S is true in all possible worlds considered as actual.

In the previous chapter we saw that if we apply this claim (2D) to a posteriori conditionals such as P→Q it follows that these conditionals are false at some possible world (considered as actual). What would happen if we apply this model to a posteriori identity statements such as (A) and (C) above? We have seen that the Distinct Property Model does not have clear anti-physicalist consequences, but could we offer a new version of the argument in terms of (2D)?

[71] See White (forthcoming).
[72] Chalmers (forthcoming: 34-35) also reconstructs the property dualism argument as a version of the two-dimensional argument.

Let's recall here the a posteriori token-identity statement (A):

(A): Smith's pain at *t* is Smith's C-fibre firing at *t*

If we assume the (2D) principle, then it follows that there is a possible world (considered as actual) where (A) is false. In addition, if we apply (2D) to type-identity statements, such as (C): 'being hurtful is being a neurophysiological state of type N', then it follows that (C) is false at some possible world considered as actual. What are the consequences of this? Is this a problem for the alleged truth of such identities? The following argument makes clear that it does pose a problem:

The Two-Dimensional Argument against Type-identities

I. The identity-statement 'Pain is C-fibre firing' is a posteriori.
II. If an identity statement 'A=B' is a posteriori, then the terms 'A' and 'B' have different primary intensions.
III. Therefore, 'pain' and 'C-fibre firing' have different primary intensions.
IV. The primary and secondary intensions of 'pain' coincide.
V. The primary and secondary intensions of 'C-fibre firing' coincide.
VI. Therefore, 'pain' and 'C-fibre firing' have different secondary intensions.
VII. If 'pain' and 'C-fibre firing' have different secondary intensions, then 'pain is C-fibre firing' is false.

———————

VIII. 'Pain is C-fibre firing' is false.

Let's examine each premise in turn. Premise I seems obvious. Premise II is an application of the two-dimensional model to the case of identity statements. (2D) says that if an identity statement is a posteriori, then there is a possible world considered as actual where it is false. That is, there is a possible world (considered as actual) where the two terms of the identity statement refer to different entities. That is to say, the primary intensions of such terms are different, since they have different extensions in some possible world considered as actual. So we can see how (2D) gives support to premise II. This premise is similar to the model for explaining a posteriori identities that we have explained above, namely, the Distinct Property Model (DPM). We can

call the new model stated in II the '*Distinct Primary Intension Model*' or (DPIM).[73] Put in these terms, we can see how this version of the two-dimensional argument can be seen as a reconstruction of the Property Dualism Argument above.

It is clear that III follows from I and II. Premise IV relies on the assumption that for phenomenal terms, the primary and secondary intensions are the same. Likewise for physical terms such as 'C-fibre firing', according to premise V.[74] From III, IV and V we can infer that 'pain' and 'C-fibre firing' have different secondary intensions. Why is this a problem for the truth of the identity statement 'pain is C-fibre firing'? Well, it seems plausible to say that both 'pain' and 'C-fibre firing' are rigid designators, that is, they refer to the same entity (a property or a kind, in this case) at every possible world (considered as counterfactual). Therefore, if 'pain' and 'C-fibre firing' have different secondary intensions (that is, if they have different referents at some possible world considered as counterfactual), they will also have different referents at the actual world, since each term has the same referent at all possible worlds. So if 'pain' and 'C-fibre firing' differ in secondary intensions, they will differ with respect to their referents in the actual word. That is, they will refer to different properties, and therefore the identity statement will be false at the actual world. This is why VIII follows from VI.[75]

The argument concludes that the type identity 'C-fibre firing is pain' is false. *Mutatis mutandis*, we could formulate versions of the argument involving other type-identities, such as (C) above, or token-identities, such as (A) above. So we can formulate versions of the two-dimensional argument against any a posteriori identity statement involving phenomenal and physical types or tokens.

The crucial premise in these arguments is precisely premise II, that is, the Distinct Primary Intension Model (DPIM), which follows directly from (2D).

---

[73] See Chalmers (forthcoming: section 11.2), for a compelling discussion about the differences between (DPM) and (DPIM), and in particular, about the differences between *properties* and *primary intensions*.

[74] One crucial assumption here is that terms such as 'C-fibre firing', *qua* physical terms (i.e. those appearing in a complete physical description of the world), have the same primary and secondary intensions. The other crucial assumption is that for phenomenal terms, the primary and secondary intensions also coincide. See discussion in 2.6 above, where I offered some discussion of these assumptions and I also explored alternative versions of the two-dimensional argument that did not rely on them. For the sake of simplicity, I will just make these assumptions here.

[75] Compare the identity statement 'pain is C-fibre firing', which is necessary if true, with another identity statement: "the first Postmaster General is the inventor of bifocals" (Kripke (1980: 329)). In the second case, for each of these two expressions, the primary and secondary intensions coincide. However, they are not rigid designators, so even if the terms are co-referential in the actual world (both referring to Benjamin Franklin), they could refer to different people at some other possible world.

Therefore, criticising (DPIM) is another way of criticising (2D), which is crucial for the conceivability argument against physicalism. In the following chapters, we will examine this model in more detail, and we will see what is wrong with it. But before doing that, we will examine two other epistemic arguments that are closely related to the two-dimensional argument, and which can also be seen as precursors of it, namely, Kripke's Modal Argument and Jackson's Knowledge Argument.

3.2. The Modal Argument

This argument was introduced by Saul Kripke in his *Naming and Necessity* (1972/ 1980). The Modal Argument is directed against identity theses involving phenomenal and physical properties, and in particular against two versions of the identity thesis: token-identity statements, such as "Jones's pain at 06.00 was his C-fibre stimulation at that time" (Kripke (1980: 329)) and type-identity statements, such as "pain is the stimulation of C-fibres" (1980: 329).

We will mostly focus on the version of the argument against type-identity theses. Kripke begins his argument by noticing that identity claims of the sort suggested by type-identity theories, such as 'Pain is C-fibre firing', seem to be contingent. That is, we seem to be able to imagine scenarios where they are false: "it would seem that it is at least logically possible that … Jones's brain could have been in exactly that state at the time in question without Jones feeling any pain at all" (Kripke 1980: 329). In this respect, these identity claims are analogous to other theoretical identifications, such as the familiar (1): 'Water is H_2O' and (2): 'Heat is molecular motion', since (1) and (2) also seem to be contingent. However, Kripke argues that this is an *illusion*: if water is identical to H_2O, then they are identical in every possible world. This claim can be further motivated by focusing on the identity statements (1) and (2): the terms involved are *rigid designators* (that is, terms that refer to the same entity in all possible worlds), or so Kripke argues. And identity statements involving rigid designators will be necessary if true. Therefore, if (1) and (2) are true, they will be necessarily true.

However, as we have pointed out, these identity statements seem to be contingent. Kripke argues that this appearance of contingency can be explained away. In particular, he puts forward a model for explaining why identity statements of that sort appear to be contingent even if they are indeed necessary. Nonetheless, argues

Kripke, this strategy for explaining away the appearance of contingency in the case of those identity statements cannot be applied to the case of identities concerning conscious mental states such as pain. We will first explore what that strategy amounts to, and we will then see why Kripke thinks that it cannot explain away the appearance of contingency of sentences involving phenomenal terms such as 'pain is C-fibre firing'.

Kripke's strategy for explaining away the illusion of contingency goes as follows. Let's consider statement (2):

(2) Heat is molecular motion

(2) is an a posteriori truth, discovered by scientists. Both 'heat' and 'molecular motion' are rigid designators, therefore (2) is a necessary truth. But it seems contingent: we can imagine worlds where heat is not molecular motion. Can we explain away this appearance of contingency?

Kripke explains how: "the strategy was to argue that 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" (Kripke (1980: 331)). That is, the identity statement (2) seems contingent because we are able to imagine that we are in the same epistemic situation as we are now (with respect to such statement), with such statement being false. In particular, "someone could have sensed a phenomenon in the same way we sense heat, that is, feels it by means of its production of the sensation we call 'the sensation of heat' (call it '*S*'), even though that phenomenon was not molecular motion" (331). In this case, the epistemic situation is relevantly similar because there is a phenomenon that produces in us the sensation *S*, but such a phenomenon is not molecular motion. This situation is perfectly possible. This can explain, according to Kripke, why we tend to think that (2) could be false: we imagine a situation in which the phenomenon that produces the sensation of heat is not molecular motion, and we misdescribe this situation as one in which heat itself is not molecular motion.

Therefore, Kripke suggests that we can explain away the illusion of contingency of some a posteriori identities by focusing on some qualitatively analogous statement, which is indeed contingent, and which we confuse with the original one. However, the notion of a qualitatively analogous statement is not

completely clear. The only elucidation that Kripke offers is by way of suggesting a "simpler strategy" (331) for explaining away the illusion of contingency,[76] which can be put as follows:

We have said that 'heat' is a rigid designator. However, the description we use to fix the reference of 'heat', namely, something like $D_{h:}$ 'the phenomenon that produces the sensation of heat', is not a rigid designator. This can help us explain away the appearance of contingency of (2). When we imagine that there is a possible world where (2) is false, what we are really imagining is a possible world where (2*) is false:

(2*) The phenomenon that produces the sensation of heat is molecular motion

And (2*) is genuinely contingent: there are possible worlds where molecular motion does not produce sensations of heat, because there are no living beings to experience these sensations, or because they have a different sensibility. Furthermore, there are possible worlds where the sensations of heat are not produced by molecular motion but by another phenomenon. In any case, it is clear that these are genuine metaphysical possibilities.

So the general formula for explaining away the illusion of contingency of necessary a posteriori truths is the following: When a necessary a posteriori identity statement 'X=Y' *seems* contingent, this is due to the fact that we confuse this statement with a related one, namely, 'X*=Y*', which *is* indeed contingent (where either 'X*' or 'Y*', or both, are the non-rigid descriptions used for fixing the reference of the original rigid terms).

Kripke argues that this general strategy for explaining away the appearance of contingency in necessary a posteriori truths concerning natural kinds cannot be applied to the case of psychophysical identities such as (3):

(3) Pain is C-fibre firing

The claim is that (3) seems contingent. Since, Kripke argues, this appearance of contingency cannot be explained away, then (3) must be really contingent. But if (3)

[76] It is not clear to me whether this new strategy is offered as an elucidation of the former, or as a new strategy.

was true, it would be necessarily true. Since it is not necessarily true, it is false. And this argument can be applied to any identity statement involving a phenomenal term and a physical term, since plausibly they will be a posteriori and therefore they will appear to be contingent. Therefore, the modal argument can pose problems for all type-identity claims. [77]

The crucial question is now as follows: why can we not apply the aforementioned strategy for explaining away the illusion of contingency to the case of (3)? Kripke's line of reasoning is that 'pain' does not get its referent fixed by means of a contingent property. There is no accidental property of pain which fixes the referent of 'pain', such that pains have that property in the actual world but not in other possible worlds. We cannot use this explanation here because 'pain' does not get its referent fixed in this way. There is no non-rigid description that we could substitute for 'pain' in (3) so that we obtain a sentence that is really contingent.

One alleged candidate could be, by analogy with the 'heat' case, the description $D_{p:}$ 'the phenomenon that produces the feeling of pain'. The problem is that this description does not pick out a contingent property of pain. It is necessary for pain to produce the feeling of pain, or in other words, to feel like pain. If something satisfies D_p, then that state is a pain state. We could imagine things that satisfy the description D_h but are not heat, such as alternative phenomena that produce sensations of heat. But we cannot imagine phenomena that feel like pain and are not pain, since if something feels like pain, it is pain.

So when we imagine a possible world where instances of C-fibre firing are not accompanied by the feeling of pain, we cannot say that we mistook this possibility for a different one. Therefore, we cannot explain away the illusion of contingency of (3). Therefore, this statement is not necessary, and consequently, by the principle of the necessity of true identity statements involving rigid designators, it is not true either.

We can then reconstruct Kripke's modal argument as follows:

[77] However, here we could run into similar difficulties to those explained with respect to the property dualism argument, due to the use of the notion 'physical term'. It could be argued that if we appeal to $physical_b$ terms, then we cannot be sure that all identity statements involving a phenomenal term and a physical term will be a posteriori (since there could be phenomenal terms that turned out to be $physical_b$ terms after all, by referring to a $physical_b$ property), but on the other hand, if we appeal to $physical_n$ terms, then the claim that identity statements between physical terms and phenomenal terms are not true is no longer worrisome. Therefore, it seems safer to focus on versions of the modal argument addressed against particular psycho-physical identities.

The Modal Argument

MA1: When an a posteriori necessary identity statement 'X=Y' *seems* contingent, this is due to the fact that we confuse this statement with a related one, namely, 'X*=Y*', which *is* indeed contingent (where either 'X*' or 'Y*', or both, are the non-rigid descriptions used for fixing the reference of the original rigid terms).

MA2: The a posteriori identity statement (3): 'Pain is C-fire firing' seems contingent.

MA3: There is no statement appropriately related to (3) that could be invoked in order to explain away the appearance of contingency of (3), since 'pain' does not get its referent fixed by a non-rigid description (nor does 'C-fibre firing').

MA4: Therefore, 'Pain is C-fibre firing' is not necessary.

MA5: Since 'pain' and 'C-fibre firing' are rigid designators, if (3) is true it is necessary.

_____

MA6: Therefore, (3): 'Pain is C-fibre firing' is false.

Now, we can see that this model for explaining a posteriori identities invoked by Kripke (MA1) is pretty similar to the two-dimensional framework that we explored in Ch. 2, and in particular to the corresponding model for explaining a posteriori identities. According to the two-dimensional model, as we have seen, an a posteriori necessity is such that the *primary intension* is false at some possible worlds (which explains its *a posteriori* status), but the secondary intension is true at all possible worlds (which explains its *necessary* status). And according to Kripke, standard cases of a posteriori necessities 'X is Y' are such that, although they are necessary, the related statements 'X* is Y*' are contingent.

We can capture the core ideas of Krikpe's model in terms of the two-dimensional framework. In particular, we can explain Kripke's (somehow imprecise) idea of a necessary statement *being related to* a contingent one, in terms of the two-dimensional framework: we can say that the original statement has a primary intension that is contingent, and a secondary intension that is necessary. Let me explain this a little.

For instance, let's consider again the sentence (2): 'heat is molecular motion'. According to Kripke's view, (2) seems contingent because it is related to (2*). What does it mean to say that they are "related"? The two-dimensional framework can offer an elucidation of this idea: we can say that statement (2*) corresponds to the original statement's *primary intension*. This sounds plausible, because (2)'s primary intension and (2*) are true at exactly the same possible worlds: (2)'s primary intension would be true at those worlds where the phenomenon that produces sensations of heat is molecular motion, and likewise for (2*) above. Therefore the two-dimensional framework can be seen as a further elucidation of Kripke's model.[78] In particular, the two-dimensional explanation of the posteriori necessity (2) would have it that (2) seems contingent because it is associated with a contingent primary intension, which would be true at those possible worlds where (2*) is true, namely, those possible worlds where the phenomenon that causes sensations of heat is molecular motion, and false otherwise.

The moral to be drawn from this discussion of Kripke's modal argument, for our purposes, is double: first, there is the exegetical point that some of the ideas behind the two-dimensional framework can already be found in Kripke's account of a posteriori identities. And secondly, we can see, as we also saw in our discussion of the property dualism argument above, that the two-dimensional model can bring problems not only for claims of supervenience of phenomenal truths on the physical, such as the statement 'If P then Q', but also for claims concerning type-identities and token-identities involving phenomenal properties and physical properties. It will be interesting to compare these two sorts of arguments and see how they are affected by the different responses to conceivability arguments. We will be concerned with these issues in the following chapters. Before that, we will discuss another influential epistemic argument against physicalism, namely, the knowledge argument.

3.3. The Knowledge Argument

[78] See Chalmers (1996: Ch. 3), and Chalmers (forthcoming: section 11.4), for further comparison between Kripke's modal argument and the two-dimensional argument. Chalmers (forthcoming) also offers a detailed reconstruction of Kripke's notion of 'qualitatively identical epistemic situations', in terms of what he calls 'Kripkean intensions' of a statement, that is, functions from possible worlds whose individual at the centre is in the same epistemic situation and making a similar statement, to the truth-value of such statement in those worlds. Chalmers argues that this model has some disadvantages with respect to the standard two-dimensional model.

Jackson's knowledge argument[79] introduced the now popular imaginary character of Mary, a colour-vision scientist who had spent all her life in a black-and-white room, and who learned all the physical facts about colour vision and the brain by means of a black-and-white television. One day, though, she was released from her black-and-white room, and saw a red rose for the first time. This raises the question: does she learn something new, when she sees a red rose for the first time? It seems very plausible to say that she does learn some new fact about colour vision, namely, what it is like for subjects like us to see red. Before being released from her room, she knew all the physical facts about brains and colour-vision, and in particular, Mary had access to all the relevant physical information about the brain states responsible for our red-experiences. However, Jackson argues that she could not come to know what it is like to undergo the conscious mental state of seeing red: she had to have the experience of seeing red, in order to know what it is like to have such experience.

This shows, according to Jackson, that physical information about the world leaves something out, namely, information about phenomenal properties. Jackson infers from this that a physical description of the world is incomplete, and therefore there is something over and above the physical world. Therefore, physicalism is false.

This, in a nutshell, is the so-called Knowledge Argument against physicalism,. It has received several formulations and has generated a lot of discussion, which we cannot discuss here in detail.[80] But I would like to discuss at least two of the main formulations of the argument.[81]

On a first approximation, we can formulate the knowledge argument as follows:

The Knowledge Argument

KA1: Mary knows all the physical facts.

KA2: Mary does not know all the phenomenal facts.

_____

KA3: Some phenomenal facts are not physical facts.

[79] First presented in Jackson (1982) and further developed in Jackson (1986). Precursors of this argument are in Nagel (1974) and elsewhere.

[80] See Ludlow, Nagasawa and Stoljar (2004), for a very interesting collection of papers discussing the knowledge argument.

[81] In what follows, I will draw mainly from the discussion in Chalmers (2003) and (forthcoming).

There are many criticisms of this formulation of the argument. One major problem (suggested by Chalmers (forthcoming)) appeals to the two notions of the physical that we introduced in Ch. 1: 'physical' in the narrow sense ($physical_n$), i.e. physical entities and properties of the kind that are posited by physical theory, and 'physical' in a broad sense ($physical_b$), that is, properties that globally supervene on the $physical_n$. Chalmers argues that the best way of understanding KA1 is by invoking the narrow notion of 'physical': it seems plausible to say that Mary knows all the physical facts that have been postulated in the best physical theories about colour-vision, that is, the $physical_n$ facts. We cannot assume that Mary knows all the $physical_b$ facts, because we have granted that she does not know the phenomenal facts, but we cannot just assume that these facts are not $physical_b$ facts, since this would beg the question against the physicalist. On the other hand, if we understand KA1 as invoking only $physical_n$ facts, then the conclusion should also invoke $physical_n$ facts, as follows:

KA1': Mary knows all the $physical_n$ facts.
KA2: Mary does not know all the phenomenal facts.
_____
KA3': Some phenomenal facts are not $physical_n$ facts.

This version of the knowledge argument does not commit equivocation on the notion of 'physical', but the problem now is that the conclusion KA3' does not pose any problem for physicalism: as we explained earlier, what physicalism entails is merely that all facts are $physical_b$ facts.

Chalmers (forthcoming) advances another formulation of the knowledge argument, which does not seem to suffer this problem, and which, he claims, is closer to the spirit of Jackson's original formulation. I will call this the deducibility version of the knowledge argument, or the *deducibility argument*, for short. The main idea here is that although Mary knows all the $physical_n$ truths about colour vision and the brain, and she is able to deduce all the truths that follow a priori from that $physical_n$ information, she is still not able to know what it is like to see red, for example. According to this version of the knowledge argument, this lack of deducibility of phenomenal truths from $physical_n$ information shows that physicalism is false. More precisely, the argument can be put as follows:

The Deducibility Argument

DA1: Mary can know all truths that are deducible from the $physical_n$ truths about the world.

DA2: Mary cannot know all the phenomenal truths.

DA3: If there are truths about the world that are not deducible from the $physical_n$ truths about the world, then physicalism is false.

_____

DA4: Physicalism is false.

This argument is more promising than the previous version, since it is clear that there is no equivocation on the notion of 'physical'. The crucial premise here, of course, is DA3. This premise has the typical form of the epistemic arguments against physicalism. First, an epistemic gap between the physical and the phenomenal is posited (in this case, the fact that we cannot deduce phenomenal truths from a complete $physical_n$ description of the world). And, secondly, it is inferred from this epistemic gap that there is an ontological gap (namely, that phenomenal properties are not $physical_b$ properties and therefore physicalism is false).

The main task for the advocates of the knowledge argument is to provide a satisfactory defence of DA3 above, that is, of such inference from the lack of deducibility to the falsity of physicalism. It is not an easy task to motivate such an inference. Therefore, the knowledge argument still remains a controversial argument.

As we saw in the previous chapter, one of the advantages of the conceivability argument over other sorts of epistemic arguments is that the former makes more explicit how to move from epistemic premises to ontological conclusions, namely, by means of a *modal* premise. The conceivability argument goes from an epistemic premise about what is conceivable, to a modal premise about what is possible, and from this, it moves to an ontological conclusion about what the actual world is like.

In any case, it is also possible to see the knowledge argument itself as a form of conceivability argument. In particular, we might try to defend DA3 (i.e. the inference from the lack of deducibility of phenomenal truths from $physical_n$ truths, to the failure of physicalism) by means of a *modal* claim. Chalmers (forthcoming) has presented a reconstruction of the knowledge argument along these lines. We can put this formulation of the argument as follows: Again, let P be a complete $physical_n$ description of the world (including T, the "that's all" clause), and let Q be any

phenomenal truth (for instance, that a certain subject is having a red-experience at time *t*). According to Jackson's thought experiment, it seems plausible to say that Mary could not infer a priori Q from P, when she was in her black-and-white room, even if she had no limitations in her reasoning abilities and her knowledge of the $physical_n$ truths. This suggests that the material conditional 'If P, then Q' (P→Q) is not a priori true.[82] Then, the argument would go like this:

The Knowledge Argument (2D version)

KA1*: P→Q is not a priori true.

KA2*: If P→Q is not a priori true, then it is not primarily necessary.

KA3*: If P→Q is not primarily necessary, then it is not secondarily necessary.

KA4*: If P→Q is not secondarily necessary, then physicalism is false.

———————

KA5*: Physicalism is false.

Here, the first premise is a combination of premises DA1 and DA2 above, and it is motivated by reflection on the thought-experiments about Mary (and Experienced Mary). The other three premises are a defence of premise DA3 above, based on the two-dimensional framework that we saw in the previous chapter. Remember the core claim of the two-dimensional framework:

[82] As Chalmers suggests, there is a loophole in this step of the argument. The fact that Mary could not infer Q a priori from P does not really entail that 'If P, then Q' is not a priori true. For Mary lacked some crucial concepts involved in such conditional, namely, the phenomenal concepts involved in Q, such as the phenomenal concept 'experience of red'. Arguably, it is necessary to have had the experience of red in order to possess the phenomenal concept 'red-experience'. And what is at issue here is whether a subject who possesses *all* the constituent concepts is in a position to know that 'If P, then Q' is a priori true. So Mary is not a relevant subject. However, we can solve this problem by slightly modifying the thought experiment. We can focus on Experienced Mary, who is very similar to Mary, the only difference being that one day, while she was still at her black-and-white lab, Experienced Mary was offered several coloured cards, with different sample colours: red, blue, green, and so on. She was finally able to learn what seeing these colours is like, but she could not associate each colour with the corresponding word. She now possesses the corresponding phenomenal concepts, but it still seems plausible to say that she cannot infer some phenomenal truths a priori from P. For instance, imagine that she has a complete physical description of a subject, Lucy, who is looking at a red rose. Experienced Mary knows that Lucy is having what we call 'a red-experience', and she has had this sort of experience herself, but she still cannot know what it is like for Lucy to see a red rose, because she cannot know which one of the colour-experiences she had is associated with the expression 'red-experience'. So this suggests that P→Q is not a priori, because Experienced Mary possesses all the relevant concepts but she is not in a position to know that such conditional is true. (Stoljar (2005b) introduces a similar example.)

(2D) A sentence S is a priori if and only if S is true in all possible worlds considered as actual.

It follows from 2D that if a sentence is not a priori, then the sentence is not 1-necessary (that is to say, the sentence is false at some possible world considered as actual). Therefore, if P→Q is not a priori true, it follows that it is not 1-necessary. And then, if we assume that the primary intension and the secondary intension of the sentence P→Q coincide, we can infer that P→Q is not 2-necessary either. That is to say, if P→Q is false at some possible world considered as actual, then it will be false at some possible world considered as counterfactual.[83] This entails that physicalism is false. So under this reconstruction, the knowledge argument is just another version of the conceivability argument, and both will stand or fall together.

Thus, we can reconstruct the knowledge argument as a version of the two-dimensional argument against physicalism explained in the previous chapter, and in this way, we can overcome an important problem faced by standard formulations of the knowledge argument, namely, the alleged equivocation on the notion of 'physical'. Similarly, in section 3.1, we have seen that the property dualism argument faces a similar problem: when it is intended as an argument against physicalism, and in particular, as an argument for the conclusion that phenomenal properties are not *physical* properties, there seems to be no single notion of 'physical' that makes the argument compelling.[84] We have offered a new version of the two-dimensional argument, namely, the two-dimensional argument against *type-identities*, which seems to capture the core ideas behind the property dualism argument and the modal argument, and which does not seem to commit equivocation on the notion of 'physical'.

Therefore, after presenting and assessing different epistemic arguments against physicalism in great detail in this chapter and the previous one, we have arrived at two arguments that seem to be initially compelling, namely, the two-dimensional argument against physicalism and the two-dimensional argument against type-identities. In the rest of the thesis, we will be mainly concerned with criticising

[83] Unless, that is, we endorse some version of type-F monism. If so, this inference will not be accepted. See Ch. 2 for further discussion on this.
[84] I also suggested that a similar problem could affect the modal argument, when this is intended as general argument against any psychophysical identity. See footnote 77 above.

these arguments and, especially, the core claim of two-dimensional semantics (2D) which supports both of them.

In the following (and final) section, I will introduce two of the main physicalist responses to two-dimensional arguments, namely, type-A physicalism and type-B physicalism. After briefly discussing the type-A response, we will focus on type-B physicalism, which will be our main focus in the following chapters.

3.4. Type-A vs. Type-B Materialism

In this section, we will distinguish two important materialist responses against conceivability arguments. As I emphasised earlier, we will focus on how they respond to the two-dimensional versions of the conceivability argument. Let's recall here the two-dimensional argument against physicalism:

> *The Two-Dimensional Argument (against physicalism)*
> 2DA1: P&~Q is conceivable.
> (CP): If sentence S is (ideally) conceivable, then S is primarily possible.
> 2DA2: If P&~Q is conceivable, P&~Q is primarily possible.
> 2DA3: If P&~Q is primarily possible, P&~Q is secondarily possible.
> 2DA4: If P&~Q is secondarily possible, Physicalism is false.
> _____
> 2DA5: Physicalism is false.

As we can see, in this formulation of the argument we have made explicit the premise which warrants the move from 2DA1 to 2DA2, namely, the conceivability-possibility link (CP), as we saw in Ch. 2. With this formulation in mind, we can introduce the main responses to this argument.

An important response to the two-dimensional argument is to deny the conceivability of P&~Q. This view has been labelled *type-A physicalism* (or materialism), and it typically denies that zombie worlds and inverted spectrum-worlds are (ideally) conceivable. Type-A physicalism claims that the material conditional 'If P, then Q' is ultimately a priori true. That is, P&~Q is not (ideally primarily) conceivable, neither positively nor negatively conceivable.

Another prominent response to the two-dimensional argument, so-called *type-B physicalism* (or materialism), rejects its second premise (2DA2), that is, they reject the idea that the conceivability of P&~Q entails the 1-possibility of P&~Q.[85] They reject this premise because they reject the principle that gives support to it, namely, the conceivability-possibility link (CP). Recall that there are two different formulations of this principle:

(CP+) Ideal positive conceivability entails primary possibility.
(CP-) Ideal negative conceivability entails primary possibility.

These two principles in turn give rise to two different versions of the two-dimensional argument, as we saw in Ch. 2. Hence, the advocates of the type-B response to conceivability arguments deny both principles (CP-) and (CP+), and consequently, they claim that the corresponding versions of the two-dimensional argument are both wrong.

The main idea behind this response is that the *epistemic gap* between physical and phenomenal truths, that is, the fact that physical truths do not entail phenomenal truths a priori, does not entail an *ontological gap* between physical and phenomenal facts. With respect to the conceivability argument, type-B physicalism asserts that the conceivability of P&~Q does not entail that P&~Q has to be true at some possible world considered as actual. And this is so because the conceivability-possibility thesis (CP) is false, and therefore, the fact that a sentence such as P&~Q is conceivable does not imply that it is 1-possible.

Type-B physicalism is committed to the truth of Physicalism, as characterized in Ch. 1. That is, type-B physicalism is committed to P&~Q being impossible. (In particular, Physicalism is committed to P&~Q being 2-impossible. As we saw in Ch. 2, it is plausible to assume that if P&~Q is 1-possible, then it is 2-possible too.[86] Therefore, it follows that if P&~Q is 2-impossible, it will be 1-impossible too. Then, it is plausible to say that type-B physicalism is committed to P&~Q being 1-impossible). Type-B materialists also accept the existence of an epistemic gap

[85] These labels (type-A vs. type-B physicalism) are introduced in Chalmers (1996: 165-8), and used elsewhere.
[86] As we also explained in Ch. 2, the view known as Russellian monism disputes this inference, but we can put it aside for our purposes here.

between P and Q. Hence, Type-B materialism is committed to the following three claims:

(a) (CP) is false, that is, the conceivability of S does not entail its primary possibility.
(b) P&~Q is conceivable.
(c) P&~Q is not 1-possible.

On the other hand, type-A materialists typically accept the second premise of the two-dimensional argument, that is, they do think that if P&~Q were conceivable, it would therefore be possible. But they crucially reject the antecedent of such a conditional. Therefore, the main difference between Type-A and Type-B physicalism amounts to this: whereas type-A physicalists deny 2DA1 and endorse 2DA2 in the argument above, type-B physicalists endorse 2DA1 but deny 2DA2. Or in other words, whereas type-B physicalists endorse both (a) and (b), type-A physicalists deny both (a) and (b). But they both agree that (c) is correct: P&~Q is not 1-possible. Therefore, they are both physicalists.

We will begin our exploration of these responses by looking at premise 2DA1 (the conceivability of P&~Q), and in the following chapters we will turn to premise 2DA2 and the principle (CP) that supports it, which will be our main concern.

Let's start with the question: Is P&~Q (ideally) *negatively conceivable*? Remember that a sentence is negatively conceivable when it cannot be ruled out a priori. The crucial question then is this: does P a priori entail Q? If it does not, then P&~Q is not a priori false, so it will be negatively conceivable. Many philosophers have argued that in effect P does not entail Q a priori. The idea here is that if we seem able to imagine coherently physical duplicates of ourselves without phenomenal consciousness, then this means that P does not entail Q a priori, and therefore P&~Q is conceivable.[87]

However, as we have seen, not everyone is convinced by these arguments: the so-called 'type-A physicalists' deny that P&~Q is *ideally* (negatively) conceivable. What sort of reasons do they provide in support of this? Many of them endorse a

[87] Chalmers offers a detailed defence of the conceivability of zombies and zombie worlds, and more generally, the conceivability of P&~Q, in his (1996: Ch. 3) and (2003), and elsewhere. Other advocates of the conceivability of P&~Q are Levine (2001), Loar (1999) and many others.

version of what is known as *analytic functionalism*.[88] This view proposes a certain analysis of *phenomenal concepts*, that is, the concepts that we use to refer to phenomenal states and properties, by virtue of *what they are like* to us. According to this view, phenomenal concepts can in principle be analysed *a priori* in functional terms (that is, in terms of the functional roles of the different phenomenal properties). Therefore, given that a physical description of the world (P) would entail what functional roles are fulfilled, P can also entail what phenomenal concepts are instantiated. In this way, P can entail phenomenal truths a priori, such as 'Lucy is having a headache at 2pm'.

According to analytic functionalism, phenomenal truths, like most empirical truths couched in ordinary language, can be given a reductive explanation in physical terms. Reductive explanations of this sort fall under the model that I will call '*functional reduction'*. The model works like this: for any ordinary concept C (which refers to property F), a reductive explanation of truths involving C would require the following two steps:

> *A Priori Stage*: First, we offer an a priori analysis of concept C in causal-functional terms, that is, in terms of F's typical causes and effects (or more generally, F-role).
> *Empirical Stage*: Second, we find out what physical properties realize the F-role.[89]

To illustrate, let us first examine the case of a non-phenomenal, ordinary truth, such as a truth about water. The idea here is a familiar one: first, we can define the concept WATER in the following terms:

> (WATER): If x is the watery stuff (the unique odourless, colourless liquid of our acquaintance that falls from the sky, fills rivers and lakes and so on), then x falls under WATER.

[88] Some proponents of this view are Jack Smart, Janet Levin, Frank Jackson and David Lewis, among others. See, for instance, Smart (1959), Levin (1991), Lewis (1994) and Jackson (2003).
[89] This model is presented in Levine (1993) and elsewhere, and discussed in Block & Stalnaker (1999), among others.

The functional reduction of water-truths proceeds as follows. According to the first stage of the reduction, something falls under the concept if it satisfies a certain causal-functional role, namely, the "watery stuff" role. Hence, if we find out what stuff satisfies such a role, we can infer that that stuff is water.

According to the advocates of this model of functional reduction, for every case of successful functional reduction of property F, we can show that there is an a priori entailment from P to truths about F-facts. Let's see how this works for the case of water: I will offer an a priori reasoning whose first premise is about certain microphysical facts, namely that 60% of the Earth is covered by H_2O. The conclusion of this reasoning will be an ordinary truth about water:

(i) 60% of the Earth's surface is covered by H_2O. (Physical truth)

(ii) If x is watery stuff (the unique odourless, colourless liquid of our acquaintance that falls from the sky, fills rivers and lakes and so on), then x falls under WATER. (A priori stage)

(iii) H_2O is the unique odourless, colourless liquid of our acquaintance that falls from the sky, fills rivers and lakes and so on. (Empirical stage)

(iv) (Conclusion): 60% of the Earth's surface is covered by water.

According to this argument, it is possible to a priori infer the truth about water (iv) from physical facts (i) and (iii). The second premise is supposed to be a conceptual truth, therefore we can go from (i) and (iii) to (iv) a priori.

The two steps of the functional reduction enable us to see how certain macroscopic properties are realised on a lower level, usually the physical level. Once we have a functional reduction of this kind, we can a priori infer macroscopic truths about the instantiation of the relevant property from microphysical truths, because we are able to see that the microphysical facts realize a certain causal-functional role, which entails a priori that a certain macroscopic property has been instantiated. This is so because of the concept C that we use to refer to that property: to fall under C is precisely to realize such and such functional description. Therefore, functional explanations of that sort lead to a priori entailment of the kind proposed. [90]

[90] Advocates of this model are: Levine (1993), Chalmers (1996), Jackson (1998a) and Kim (2005), among others. All these philosophers hold that this model can be applied to ordinary macroscopic truths, such as those about water, heat, boiling, etc. However, they disagree when it comes to

Analytic functionalists claim that this model can also be applied to the case of phenomenal truths and the phenomenal concepts they involve. Let us take the example of the phenomenal concept 'headache'. The functional analysis of 'headache' could be something like this:

> (HEADACHE): If x is a state caused by such and such stimuli, and which causes such and such behavioural responses, then x falls under HEADACHE.

With this allegedly *a priori* analysis of the concept HEADACHE, we could infer our phenomenal truth 'Lucy is having a headache at 2pm' from a physical description of the world P, as follows:

(i) Lucy is having a token of C-fibre firing at 2pm. (Physical truth included in P)
(ii) If x is a state caused by such and such stimuli, and which causes such and such behavioural responses, then x falls under HEADACHE. (A priori stage)
(iii) C-fibre firing is a state caused by such and such stimuli, and which causes such and such behavioural responses. (Empirical stage)
(iv) Conclusion: Lucy is having a headache at 2pm.

This is the sort of a priori entailment which, analytic functionalists suggest, is available for all phenomenal truths. If so, then it is clear that the conditional 'If P, then Q' will be a priori true (recall that Q is any arbitrary phenomenal truth), and therefore, P&~Q will not be conceivable.

One controversial step, though, is precisely premise (ii), that is, the allegedly a priori functional analysis of HEADACHE (and mutatis mutandis, for other phenomenal concepts). The problem is, of course, that it is controversial to claim that this is an a priori truth.[91]

---

consciousness: for instance, Jackson, *qua* analytic functionalist, believes that it is possible to offer a reductive explanation along these lines for phenomenal truths, whereas Chalmers, Levine and Kim think it is not possible.

[91] See Chalmers (1996) and (2003) for further discussion of analytic functionalism, and in particular, for a critique of the proposed functional analyses of phenomenal terms.

There are other arguments against the claim that P&~Q is conceivable that do not seem to be committed to analytic functionalism. Different philosophers have suggested different reasons for which P&~Q might not be (ideally) negatively conceivable after all. Some of these proposals argue that such conceivability would entail some form of epiphenomenalism, (that is, the view that phenomenal properties are not causally efficacious with respect to physical properties).[92] Some others argue that P&~Q, even if prima facie (negatively) conceivable, is not ideally conceivable after all.[93] All these views deserve attention, but in my view, the negative conceivability of P&~Q has enough intuitive plausibility to justify the discussion of the question that we will focus on, namely: what follows if we assume that P&~Q is (ideally negatively) conceivable?

As I explained in Ch. 2, Chalmers claims that positive conceivability is probably a better guide to possibility than negative conceivability. Therefore, in order to make a stronger case against materialism he has to argue that P&~Q is not only negatively conceivable but also positively conceivable. He tries to achieve this goal by means of two strategies. On the one hand he tries to bridge the gap between negative conceivability and positive conceivability in the general case.[94] On the other hand he tries to give direct evidence for the positive conceivability of P&~Q (that is, the claim that there are scenarios whose descriptions entail P&~Q a priori).

The direct argument for the positive conceivability of P&~Q, or, more concretely, for the conceivability of zombie worlds, is mainly the claim that we can conceive of a possible scenario physically identical to this world but without phenomenal consciousness. For Chalmers this is perfectly conceivable, there is no contradiction in the description of such a scenario, and we could fill in the gaps of this rough description, without finding any contradiction. Chalmers believes that a basic description of such a possible scenario would entail P&~Q, so it is positively conceivable. Again, there are those who deny this, but it seems plausible enough to

[92] See, for instance, Kirk (2005) and Perry (2001). Two initial problems with this view are the following: first, it is not clear to me why, if the conceivability of P&~Q entailed epiphenomenalism, this would show that P&~Q is not conceivable after all. For it is not clear to me that epiphenomenalism is a priori false. In addition, it is not clear either that the conceivability of P&~Q is indeed a priori committed to epiphenomenalism. For it does not seem a priori incoherent to assume that some form of interactionism is correct, that is, that phenomenal properties are causally efficacious with respect to physical properties after all, even if they are not physical themselves. This might be false, but it does not seem *a priori* false. In any case, there is much more to be said about these questions.

[93] See, for instance, Warley (2003) or Marcus (2004).

[94] We discussed this issue in Ch. 2.

motivate our discussion here: we will henceforth assume that P&~Q is both (ideally) negatively and positively conceivable and we will examine whether this entails its 1-possibility or not.

I will argue that, *were* P&~Q to be conceivable, its possibility would not necessarily follow. I will not offer a full-fledged defence of type-B materialism, since I will not be committed to the 1-impossibility of P&~Q, but I will offer a detailed defence of a core claim, namely, that the conceivability of P&~Q does not entail its possibility. In this way, I will contribute to the defence of type-B physicalism.

4. The Non-Exceptionalist Strategy against Conceivability Arguments

4.1. Introduction

In this chapter we will focus on the critique of the two-dimensional arguments that we have presented earlier. We will introduce two important strategies against such arguments, both of which fall under the type-B response that we explained in the previous chapter. That is, both strategies hold that, even if there is an epistemic gap between physical and phenomenal truths, this does not imply an ontological gap.

Let's start with the two-dimensional argument against physicalism, which we can repeat here for easy reference:

The Two-Dimensional Argument (against physicalism)
2DA1: P&~Q is conceivable.
(CP): If sentence S is (ideally) conceivable, then S is primarily possible.
2DA2: If P&~Q is conceivable, P&~Q is primarily possible.
2DA3: If P&~Q is primarily possible, P&~Q is secondarily possible.
2DA4: If P&~Q is secondarily possible, Physicalism is false.
———————
2DA5: Physicalism is false.

Both of the strategies that I want to introduce focus on a crucial premise of this argument, namely, (CP):

(CP): If sentence S is (ideally) conceivable, then S is primarily possible.

(As we saw in Ch. 2, there are two versions of this principle, according to whether we understand 'conceivability' as negative or positive conceivability, but we can ignore this difference in what follows, since what these two strategies have to say applies to both.) This thesis (CP) clearly entails that if P&~Q is (ideally) conceivable, then it is 1-possible (that is, premise 2DA2 in the two-dimensional argument above). And as we have seen before (in Ch. 2), (CP) follows from the central claim of two-dimensionalism, namely, (2D):

(2D) A sentence S is a priori if and only if S is 1-necessary (i.e. true in all possible worlds considered as actual).

Our main aim will be to examine whether (CP) is correct. In particular, we will examine whether there are plausible counterexamples to such a claim. If we establish that there are genuine counterexamples, then we will have refuted (CP). And by doing this we will manage to do two things: first, we will show that (2D) is incorrect, since (2D) is committed to the truth of (CP). Thus, if there are counterexamples to (CP), (2D) will be false too. And in addition, if we show that (CP) is false, we can conclude that premise 2DA2 is ungrounded. In other words, the two-dimensional argument against physicalism relies on the principle (CP) to make premise 2DA2 plausible. If it is shown that (CP) is incorrect, then premise 2DA2 will remain unjustified, and therefore the argument will no longer pose a problem for physicalism, since we will have no reasons to believe that the conceivability of P&~Q entails its possibility.

4.2. Evaluating (CP): Introducing Strong Necessities

Our main task, then, is to argue that the fact (if it is a fact) that P&~Q is conceivable does not imply that it is 1-possible, that is to say, that the position that holds that P&~Q is both conceivable and 1-impossible is a coherent one. We do not have to establish that this position is in fact true, because this would amount to establishing that type-B materialism is true, and this is beyond our scope here. Our target is the following: we want to show that (CP) is false, and therefore that 2DA2 is ungrounded. This is enough to make the two-dimensional argument unpersuasive. We do not have to offer, in addition, an argument for the falsity of 2DA2.

Therefore, we will be concerned with showing that (CP) admits counterexamples. Once this has been clarified, we can turn to examining the following question: what would constitute a counterexample to (CP)? What form would they take?

A clear counterexample to (CP) is what I will call *conceivable impossibilities*, that is, sentences that are conceivable but false in all possible worlds considered as actual. We can search for cases of conceivable impossibilities by searching for the so-called *strong (a posteriori) necessities*:

Strong necessities: S is a strong necessity if and only if S is a posteriori but true at all possible worlds considered as actual (1-necessary).

We can easily see that if S is a strong necessity in this sense, then its negation ~S is a conceivable impossibility, as follows: if S is a strong a posteriori necessity, then ~S will be conceivable (it's not a priori false, since S is only a posteriori true) but false at all possible worlds considered as actual (since S is true at all possible worlds considered as actual). Hence, if there is a case of strong necessity, there is also a case of conceivable impossibility, namely, the negation of the former, and (CP) will be false.

As we witnessed in Ch. 2, it is not so easy to find examples of strong necessities. Standard cases of a posteriori necessary identities such as those discussed by Kripke (1980) and Putnam (1975) do not seem to pose any problem for (CP). To recall this point briefly, let's consider again our two examples of a posteriori necessities:

(1) Water is H_2O
(2) Heat is molecular motion

These sentences are clearly a posteriori and also true in every possible world considered as counterfactual since the terms involved are rigid designators. Let's consider now their negations:

(~1) Water is not H_2O
(~2) Heat is not molecular motion

(~1) and (~2) are not a priori false, that is, they are conceivable, but are false in every counterfactual world.

However, as we explained in Ch. 2, Chalmers and Jackson have argued that there are possible worlds considered as actual in which (~1) and (~2) are true (that is, possible worlds that *verify* those sentences). For example, (~1) is verified by Twin Earth, and (~2) is verified by a possible world where the phenomenon that causes sensations of heat is not molecular motion.

Therefore (~1) and (~2) are not cases of conceivable impossibilities, in the sense defined above, since they are true at some possible worlds considered as actual. Likewise, (1) and (2) are not cases of strong necessities, because they are not true at all possible worlds considered as actual. Hence, standard a posteriori identities do not seem to constitute counterexamples to (CP).

There is another important class of putative counterexamples to the conceivability-possibility link, apart from a posteriori identity claims, which have been discussed in the literature and have also been addressed by Chalmers and Jackson.[95] This class involves claims of necessitation between facts or truths at different levels, usually between physical and higher-level truths. For instance, let's consider again our ordinary truth about water, namely:

(W): Water covers 60% of the Earth's surface

The idea here is that the conditional P→W might be a strong necessity. That is, the suggestion here is that this conditional is true at all possible worlds considered as actual, but not a priori true. If so, (CP) would clearly be false.[96]

The issue of whether there are strong necessities of such sort is a controversial one. In the following section, we will explain two different ways of arguing that there are indeed cases of strong necessities (and therefore, conceivable impossibilities).

4.3. Exceptionalism vs. Non-Exceptionalism

Once we have explained what form counterexamples to (CP) might take, we are in a position to introduce a very useful distinction among two different types of responses to conceivability arguments: the *exceptionalist* strategy and the *non-exceptionalist* strategy (these labels were introduced in Levine (2001)). We can characterize these two strategies as two different ways of attacking the principle (CP).

The question at issue is whether there are *conceivable impossibilities*, that is, whether there are conceivable sentences that are not true at any possible world considered as actual. As we have seen, the existence of strong necessities would

[95] See, for instance, Chalmers & Jackson (2001).
[96] This sort of view is defended by Block & Stalnaker (1999) and Levine (2001), among others. We will examine these proposals in more detail later in this chapter; here I am just trying to get a flavour of what a counterexample to (CP) might look like.

imply that there are conceivable impossibilities (namely, the negations of those strong necessities). Both the exceptionalist and the non-exceptionalist strategy agree that there are strong necessities, and therefore that (CP) is false. The main difference between them, in Levine's words, is the following:

> E-types [exceptionalists] generally allow that the framework developed by the anti-materialist for handling the standard cases of a posteriori necessities is the right one. […] Their refusal to accept the anti-materialist conclusion is based on a rejection not of that framework in general, but of its application to the case of qualia. […] N-E types [non-exceptionalists] generally reject [that framework]. (Levine (2001: 50))

Thus, the *exceptionalist* strategy typically agrees that the standard examples of a posteriori necessities fit (CP). Advocates of this strategy accept the two-dimensional framework and the corresponding treatment of a posteriori necessities, and they agree that most cases fall under that model. However, they argue, when it comes to the concepts that refer to our sensations, that is, *phenomenal concepts*, (CP) and the two-dimensional framework cannot be applied. That is, sentences involving phenomenal concepts constitute counterexamples to (CP), but they are the *exception* to the rule, since (CP) is correct with respect to examples that do not involve phenomenal concepts.

On the other hand, the *non-exceptionalist* strategy argues that (CP) is incorrect even when applied to standard cases which do not involve phenomenal concepts. That is, we do not need to invoke phenomenal concepts in order to show that (CP) is wrong: advocates of the non-exceptionalist strategy claim that there are many other counterexamples to (CP), outside the phenomenal domain: sentences involving standard, non-phenomenal empirical concepts also pose counterexamples to (CP).

To illustrate this distinction, we can focus on the example we introduced above, namely, P→W. The non-exceptionalist strategy would argue that this conditional is both a posteriori and true at all possible worlds considered as actual, or in other words, that P→W is a strong necessity. On the other hand, exceptionalist responses typically have it that conditionals such as P→W fit the (CP) thesis. Therefore, P→W cannot be both a posteriori and 1-necessary. (They would argue that P→W either is a priori and 1-necessary, or a posteriori and not 1-necessary.) That is, exceptionalist responses claim that sentences not involving phenomenal terms, such

as P→W, do not constitute counterexamples to (CP). However, they think that a sentence such as the following would in effect constitute a counterexample to (CP):

(4) If P, then Lucy is having a headache at 2pm

This conditional involves the phenomenal concept HEADACHE, and since phenomenal concepts have some special features that non-phenomenal concepts lack, sentences such as (4) do not behave in the way standard sentences do. In particular, (CP) and related principles cannot be applied here. It seems plausible to say that (4) is a posteriori, but this does not entail that (4) is not 1-necessary, or so the advocates of the exceptionalist strategy argue.

This is, in a nutshell, the main distinction between exceptionalist and non-exceptionalist strategies. In the next two sections, I will explore the non-exceptionalist strategy, and I will focus on the exceptionalist strategy in chapters 5 and 6.

4.4. The Non-Exceptionalist Strategy and A Priori Entailment

4.4.1. In search of strong necessities

In this section and the next (4.5) we will be concerned with the assessment of the non-exceptionalist type of response to conceivability arguments. As we have explained in the previous section, the main idea of this strategy is that we can find many counterexamples to (CP) which do not involve phenomenal concepts, and that therefore this principle fails across the board. In what follows, we will discuss some prominent non-exceptionalist strategies in the literature such as those suggested by Block & Stalnaker (1999) and Levine (2001), among others. However, we will see that these formulations are not completely satisfactory. I will focus on a new formulation that can overcome some of the problems, and I will defend it from recent objections by Chalmers and Jackson.

The main idea behind the non-exceptionalist strategy can be put as follows. Recall the conditional that we introduced in the previous chapter, namely, P→W (that is, 'If P, then water covers 60% of the Earth's surface'). The basic idea is that conditionals of this sort constitute a counterexample to (CP). In particular, we will be concerned with conditionals about the (global) supervenience of certain macroscopic empirical facts couched in ordinary language, on the microphysical facts.

So, the non-exceptionalist strategy claims that P→W is a strong necessity, that is, P→W is both a posteriori and primarily necessary.

Let's examine the first issue first: is P→W a posteriori? Philosophers such as Block & Stalnaker (1999), Byrne (1999), Levine (2001) and others have argued that it is indeed a posteriori: they argue that we are not in a position to know conditionals such as P→W independently of experience. For instance, Levine argues that sentences such as 'P and it is not the case that water fills the lakes and oceans' are conceivable, and therefore 'If P, then water fills the lakes and oceans' is not a priori (2001: 51). Block & Stalnaker submit that we cannot infer ordinary macroscopic truths from microphysics a priori: "one cannot move a priori from microphysics alone to the conclusion that H_2O boils. […] One cannot move a priori from microphysics alone to the claim that there is water around here" (1999: 20). The strategy of these authors is the following: they argue that sentences such as P&~W are conceivable in the relevant sense but metaphysically impossible, and then the fact that P&~Q is conceivable does not imply that it is metaphysically possible.

We can grant that P→W is a posteriori. But, in order for it to be a strong necessity, it has to be true at all possible worlds considered as actual. What reasons do we have to hold that conditionals such as P→W are necessary? The main reason is that it seems very plausible that standard macroscopic facts, such as those about water, storms, tigers, gold, and so on, *globally supervene* on the physical. Then there will be conditionals of the sort 'if physical facts are such and such, then facts about water (or tigers, or gold…) will be such and such', which will be necessarily true. As we can recall from Ch. 1, the global supervenience of any property *F* on the physical was committed to any (minimal) physical duplicate of the actual world being a duplicate with respect to *F*. Therefore, for any world where the physical facts are the same as in @ (the actual world), the *F*-facts will be the same too. In particular, focusing on the macroscopic fact expressed in W, if we assume that water-facts globally supervene on the physical, then it follows that W will be true at any minimal physical duplicate of the actual world, that is to say, the conditional P→W will be necessary.

However, what this means is that such conditionals are true at all possible worlds considered as *counterfactual*, that is, P→W is 2-necessary. But in order for it to be a strong necessity, it should be also 1-necessary, that is, true in all possible worlds considered as actual. Can we assume it is?

Unfortunately, we cannot. The problem is that there might be possible worlds considered as actual where P is true but W is not. Let me explain in some detail why this is so.

The problem is this: it would be quite tendentious to claim at this stage of the discussion that P→W is 1-necessary, because if it were 1-necessary, then the conditional P→Q would have to be 1-necessary as well, at least for some phenomenal truths Q, and this is something that we cannot grant here.

Let's go slowly. The basic idea is that in order for W to hold true at possible worlds (considered as actual) where P is true, it is necessary that certain phenomenal truths hold there too. This is due to the fact that many macroscopic concepts such as WATER are such that something falls under them in virtue of, among other things, the *phenomenal experiences* caused by it. For example, when we evaluate the extension of WATER in a possible world considered as actual, we have to find out what stuff satisfies the description 'the *odourless*, *colourless* stuff of our acquaintance that fills rivers and lakes, and so on'. And, arguably, something falls under the description 'the odourless, colourless stuff' only if it is connected with the appropriate phenomenal experiences.

Let *v* be a (minimal) physical duplicate of @, and let's evaluate whether W ('Water covers 60% of the Earth's surface') is true at *v* considered as actual. What is referent of 'water' at *v*? In order to know what 'water' picks out in that possible world, we have to know what stuff is connected with certain experiences of colourless, odourless stuff, etc. in *v*. If the phenomenal properties that characterize those experiences are physical properties, as physicalism claims, then the physical duplicate *v* will be identical to @ concerning the distribution of phenomenal properties, and therefore, the same entities will fall under 'water' in the two worlds. But if, on the contrary, physicalism is false, as the conclusion of conceivability arguments has it, then we cannot be sure that our physically identical world will be phenomenally identical. Since *v* is a minimal physical duplicate, P and the "that's all" clause hold there. That is, *v* contains our physical properties and nothing else. Thus, if phenomenal properties are not physical properties, they are not instantiated at all at *v*. So we cannot be sure that H_2O is going to be causally connected to phenomenal states at *v* (since we cannot be sure there will be any phenomenal states). Then, we cannot assume that the extension of WATER in that minimal physical duplicate is H_2O. So we cannot be sure that sentence W is going to hold true there.

The conclusion is that, unless we are committed to the claim that P→Q is 1-necessary, we cannot maintain that P→W is going to be 1-necessary. But, of course, we should not be committed to the claim that P→Q is 1-necessary, since that is precisely the claim that is at stake here: we are discussing whether certain arguments against that claim are correct or not, and therefore we cannot assume that it is right.

Therefore, showing the aposteriority of P→W is not sufficient to show that P→W is a *strong necessity*, since it is controversial to maintain that P→W is 1-necessary. And since P→W can be 1-necessary only if P→Q is, we should find another example which could be 1-necessary without P→Q being 1-necessary. Obviously, we cannot presuppose at this stage that P necessitates some phenomenal truths, because this would be very close to presupposing physicalism. In my view, this is a problem for Block & Stalnaker's and Levine's responses to conceivability arguments.

Therefore, we have to be more careful in finding out what kind of examples we want to submit as counterexamples to (CP). Nonetheless, we will see that it is not very hard to modify the example of P→W so as to make it a more convincing counterexample to (CP) and related principles. Chalmers and Jackson (2001), in their response to Block & Stalnaker (1999), discuss what form a successful counterexample to (CP) might take, and they argue that there are no cases that have that form. In what follows, I will explain what these counterexamples would look like and why they ultimately fail, according to Chalmers and Jackson. I will argue that their arguments for that claim are problematic. I will conclude that we have good reasons to believe that there are counterexamples to (CP).

4.4.2. In search of a priori conditionals

Chalmers & Jackson (2001) accept the view that P→W is *not* a priori true, but they claim that we can modify it so as to obtain an a priori conditional. The modifications are as follows.

The main focus in Chalmers & Jackson (2001) is on ordinary macroscopic truths such as W and other ordinary-language truths about water, tigers, storms and so on. Let M be any macroscopic truth of this sort. The question at issue then is whether the conditional P→M is a priori or a posteriori. They argue that we have to add some information to the antecedent, P, in order to make it into an a priori conditional. The information that we have to include is the following.

First of all, recall that we are including a "that's all" clause in P (which Chalmers and Jackson label 'T'). This is needed because otherwise, P would not entail negative truths such as 'there are no angels', as we explained in Ch. 1. In what follows, then, we should bear in mind that P includes T.

Secondly, we have to add indexical information to the microphysical descriptions of a world, for example a clause of the sort 'I am here'. This is needed because otherwise we could not infer a priori a lot of information about a world from a microphysical description of it: we need to know where we are located in a world, in order to a priori deduce truths involving natural kind terms and other context-dependent terms. For instance, in order to deduce truths about water a priori from P, we have to discover what stuff satisfies the water-role, and for that we have to know where we, the speakers, are located within the microphysical description. For we have to know what stuff is the liquid of *our acquaintance*, what stuff is causally linked *to us*, and so on, because this is necessary in order for something to fall under WATER (in a given possible world considered as actual). Therefore, we need to supplement P with a clause of the form 'I am A and B is here' (I, henceforth), where A and B are microphysical characterizations of an individual and a place, respectively. In what follows, we will assume that P includes this clause containing indexical information I as well.[97]

Thirdly, we have to add phenomenal information to the microphysical description P, in order for P to a priori entail many macroscopic truths that involve or depend on phenomenal truths. This is so because there are many macroscopic concepts that depend on phenomenal truths for their application, as we have seen above with respect to the example of sentence W. To repeat the basic point: as we can witness in the case of the concept WATER, when we consider a possible world as actual, we apply the concept to some stuff on the basis of how that stuff looks to us: it has to be colourless and odourless, for instance. Otherwise it would not fall under the concept (in a possible world considered as actual). Then, in order to infer truths about water from a microphysical description of a world, we have to know what experiences subjects are having in that world. Chalmers and Jackson suggest that we can

[97] See Chalmers & Jackson (2001: 318).

supplement P with a complete phenomenal description of the world (Q_T, henceforth), in order to solve this problem.[98]

Chalmers and Jackson's final thesis about the a priori entailment of macroscopic truths is the following: for any ordinary macroscopic truth M (such as truths about water, gold, storms, and so on), the conjunction of P (which includes T and I) and Q_T entails M a priori.[99] That is, for every M, P&Q_T (or PQ_T, for short) entails a priori M.

Therefore, we can characterize the non-exceptionalist response we are evaluating as the view that the conditional $PQ_T \rightarrow M$ is *a posteriori*, at least for some M. It is clear that this conditional is 1-necessary, because for any possible world considered as actual where both P and Q_T hold, M will hold as well. Therefore, if we can show that $PQ_T \rightarrow M$ is a posteriori, we will have found a strong necessity. Chalmers and Jackson, on the contrary, argue that it is a priori. In the rest of the chapter, I will explain and evaluate Chalmers and Jackson's defence of the claim that that PQ_T entails a priori any M, and I will argue it does not work. I will conclude that we have good reasons to believe that $PQ_T \rightarrow M$ is a strong necessity, which shows that (CP) above is wrong.

4.4.3. Concepts as conditional abilities

According to Chalmers and Jackson, ordinary macroscopic truths are a priori entailed by PQ_T, and this is due to the nature of macroscopic concepts: the possession conditions of macroscopic concepts are such that subjects who possess the relevant concepts are able to infer a priori any M from PQ_T.

So, what are the possession conditions for macroscopic concepts? Chalmers and Jackson emphasise that they are not committed to the existence of explicit definitions for every concept, in the form of a set of necessary and sufficient conditions.[100] Rather, they claim, their view about a priori entailment relies on a different assumption about concepts:

---

[98] Notice that if analytic functionalism were correct, and therefore P a priori entailed Q_T, then we would not need to add Q_T to P. But, since we cannot assume that at this point, we have to explicitly add Q_T to P, in case P does not entail Q_T a priori. See Chalmers & Jackson (2001: 318-20), for further discussion.

[99] Notice that when we add Q_T to P, the "that's all" clause T changes its meaning slightly: the claim is no longer that there is nothing over and above what P says, but rather that there is nothing over and above what P *and* Q_T say.

[100] This is a response to Block & Stalnaker (1999), who criticise the possibility of providing definitions for macroscopic terms in microphysical terms.

> Once an essential role for explicit definitions is eschewed, the model of conceptual analysis that emerges is something like the following. When given sufficient information about a hypothetical scenario, subjects are frequently in a position to identify the extension of a given concept, on reflection, under the hypothesis that the scenario in question obtains. […] What emerges as a result of this process may or may not be an explicit definition, but it will at least give useful information about the features in virtue of which a concept applies to the world. […] The possibility of this sort of analysis is grounded in the following general feature of our concepts. If a subject possesses a concept and has unimpaired rational processes, then sufficient empirical information about the actual world puts a subject in a position to identify the concept's extension. (Chalmers and Jackson (2001: 322-3))

The main idea about concept-possession that Chalmers and Jackson are assuming here is that a subject possesses a concept only if she is able to find the extension of the concept in a sufficiently rich description of the world. According to Levine, this view is what divides Chalmers and Jackson from many of their non-exceptionalist opponents. Levine poses the following question: "For most terms, do we have a priori access to sufficient information to determine their referent given a context (a possible world considered as actual)?" (Levine (2001: 53)). Those who answer this question positively, such as Chalmers and Jackson, are called *ascriptivists*. Those who answer the question negatively, such as Levine himself, are called *non-ascriptivists*.

Thus, according to *ascriptivism*, when we possess a concept, say, WATER, we are able to identify the extension of the concept, given a sufficiently rich description of a possible scenario. That is, a subject who possesses a concept C will be able to identify the extension of C when presented with a scenario, and therefore, she can infer truths of the form '*x* falls under C' from a sufficiently rich description of the scenario. This means that possessors of a concept C will have the ability to infer truths involving C from some other truths, just by virtue of possessing the concept, that is, *a priori*. Ascriptivists, then, are committed to the following view:

> (**Ascriptivism**) For any concept C, there is an application conditional like this: (AC): 'If *x* is F, then *x* falls under C'.

Notice that (AC) is just a schema, where '*x* is F' stands for whatever information turns out to be required in order to find out the extension of C in a given scenario. Different

versions of ascriptivism will require different sorts of descriptions, but the core idea is that for each concept C, there are certain descriptions such that possessors of C can infer truths of the form '*x* falls under C' from those descriptions. (These descriptions can vary among subjects.)

Levine on the contrary rejects any form of ascriptivism: he claims that we can use, say, the term 'cat' so as to refer to cats, without having *a priori* access to the sophisticated information that would allow us to determine the referent of 'cat' in different scenarios. The same applies to most of our terms and concepts.

Then, according to Levine, even if we possess all the relevant concepts, we do not have a priori access to the information that would allow us to infer truths about water from a physical description. Hence, if we reject the ascriptivist idea that concepts are associated with application conditionals, it is obvious that physical truths are not going to entail truths about water a priori. That is, $PQ_T \rightarrow W$ is not going to be a priori. The non-ascriptivist response to conceivability arguments is an important one, although it is highly controversial, since it rejects the whole framework that conceivability arguments rely upon. [101] My strategy, on the other hand, is to concede as much as I can to the advocates of conceivability arguments, so that the ensuing response will be as strong as possible: if we show that conceivability arguments do not work, even on the assumption that the ascriptivist framework that they rely upon is generally correct, then this is a more damaging result for the prospects of such anti-physicalist arguments. Therefore, in what follows I will assume that the ascriptivist view is more or less in the right track, and I will explore what follows from this. I will argue that the non-exceptionalist strategy is still tenable, and therefore, I will show that ascriptivism is compatible with the existence of strong necessities and the falsehood of (CP).

Levine himself suggests that it is not necessary to reject ascriptivism in order to defend the non-exceptionalist response. As we will see, even if we accept ascriptivism, there are serious problems for the claim that $PQ_T \rightarrow M$ is a priori.

The ascriptivist view, according to which every subject associates a priori application conditionals with the concepts she possesses, is neutral concerning the

[101] There are many non-ascriptivist defences of non-exceptionalism in the literature, such as Levine (2001) himself, Block & Stalnaker (1999) and others. For instance, Block & Stalnaker say: "there is no way to fill in the details of 'the water role' so that it is a conceptual truth that water occupies the water role" (1999: 16). In the same line, Levine argues that the sentence 'water is watery' (where 'watery' is a description of the water-role) is not a priori (2001: 55-66).

information contained in the antecedents of such conditionals. That is, ascriptivism is neutral concerning what kind of information we need to know about a world considered as actual in order to apply our concepts.

Chalmers and Jackson make some more specific comments about the kind of application conditionals that, according to them, we associate with our expressions. The first positive remark they make about application conditionals is that "for many or most concepts, there will exist application conditionals (corresponding to arbitrary epistemic possibilities) whose antecedents contain nontrivially sufficient information" (2001: 325). That is, for most macroscopic concepts, there will be application conditionals of the form (AC) that do not invoke the concept C (or related concepts) in the antecedent; it will be possible to describe feature F without invoking concept C.

A second remark about application conditionals is that "it is possible that two people who use a given name might use it with different a priori application conditionals" (327). The main idea is that whilst different people could associate the same concept C with different application conditionals, it is still the case that we are able to infer truths involving C from truths that do not involve C directly. This variability does not matter for their purposes: the important issue is that, for every user of concept C, there must be some application conditional that is a priori for that subject. It is only that for some concepts these application conditionals will be subject-relative (for instance, the application conditionals for proper names and, arguably, natural kind-terms seem to be subject-relative in that sense).

From these remarks, Chalmers and Jackson conclude that for most macroscopic truths M of the form '*x* falls under C' (and for related truths), there is an application conditional E→M, where E is a non-trivial description of the world (possible or actual).

This view of concepts is more plausible than the view that concept-possession requires associating the concept with an explicit definition. The new view can solve one of the main problems of the definition-view: we usually attribute macroscopic concepts to subjects that are ignorant of, or wrong about, the corresponding definitions. Under the new view, there is no single piece of information that every subject has to know a priori in order to possess a particular concept. They just have to have a conditional ability to identify the concept's extension. They can do this in virtue of very different beliefs about the extension. But the new view is not free of commitments: according to Chalmers and Jackson, possession of a concept C cannot

happen in the absence of such a conditional ability. A subject that is not able to identify the extension of C, given enough information, does not possess the concept C.

The plausibility of this view depends on how the notion of 'enough information' is spelled out. The first condition for Chalmers and Jackson is that the information is not trivial, as we have seen before. This does not seem very implausible. The second and more crucial condition is that our concepts are associated with application conditionals such that they enable us to infer any M from PQ_T. The idea is that, since there are non-trivial application conditionals like (AC) for most macroscopic concepts, then we can use them to infer a priori macroscopic truths M (such as '*x* is C' and other related truths) from sufficiently rich descriptions, which can be ultimately inferred from PQ_T. In short: according to Chalmers and Jackson, we can find the extension of each concept in a scenario described in terms of PQ_T.

They will argue that, given that those application conditionals are part of the possession conditions of our concepts, a subject who possesses all the relevant concepts will be able to infer any M from PQ_T, and that therefore, $PQ_T \rightarrow M$ will be a priori for this subject, since she can know the truth of that conditional just by virtue of possessing the relevant concepts, without further experience.

In the next section I will spell out what specific kinds of application conditionals are needed for PQ_T to entail a priori any M. And in section 4.5, I will argue that it is simply wrong to take application conditionals of that sort to be part of the concepts' possession conditions.

4.4.4. A two-step entailment

Chalmers and Jackson argue that the a priori entailment from PQ_T to M goes through two steps. For most thinkers, it is not the case that they can directly a priori infer ordinary macroscopic truths from a description of the world in terms of PQ_T. The claim is somewhat weaker: firstly, we can move a priori from a description of the world in terms of PQ_T to macroscopic truths in the *language of physics*, that is, truths describing macroscopic objects in terms of the properties invoked in physical theory. And once we have a description of the world in macrophysical terms, then we can infer a priori all macroscopic truths about the world.

Therefore, the first step is the entailment from PQ_T to all macroscopic truths expressed in macrophysical concepts: "First, PQ_T implies complete information (in

the language of physics) about the structure, dynamics, composition, and distribution of macroscopic systems, as well as information about the actual and potential perceptual appearances that they present" (Chalmers & Jackson (2001: 329)). And the second step is that this complete macroscopic description of the world in terms of physical theory (plus Q_T) entails any ordinary macroscopic truth M.

It could seem at first sight that the information contained in Q_T is going to be very useful in order to infer a priori M. Q_T contains truths about what the thinker perceives at a given time, and from this sort of perceptual information we can usually gather what is going on around us, so it seems that we could infer many macroscopic truths from phenomenal information. The problem with this idea, as Chalmers & Jackson notice, is that this kind of inference is not a priori.[102] The inference of external macroscopic truths from phenomenal truths is a posteriori, because for any given phenomenal description it is conceivable that it is a case of misperception. That is, we can coherently imagine that our experiences are deceiving us (at least in most cases, if not all). Therefore, a given phenomenal truth such as, say, 'I have the experience of seeing a blue liquid at *t* in position *p*' is compatible with the claim that there is no blue liquid in *p* at *t*. Hence, phenomenal descriptions do not entail macroscopic truths about the external world a priori, because the former are conceptually compatible with the latter being false.

Chalmers and Jackson realise that Q_T in itself is not sufficient for ruling out alternative explanations for Q_T, but they claim that Q_T and P are sufficient for implying M a priori (that is, for ruling out alternatives to M). I will argue this is not correct.

According to Chalmers and Jackson, the information in PQ_T will rule out a priori all but one alternative, that is, the correct macrophysical description of the world. Let's see how.

As we have seen, Chalmers and Jackson claim that "a macroscopic description of the world in the language of physics is implied by a microscopic description of the world in the language of physics. Such a thesis is extremely plausible: it is not subject to any worries about translation between vocabularies, and involves only a change of scale" (2001: 330-1). This might be right for some cases: maybe if we know, for instance, the individual masses of microphysical entities x_1, x_2 … x_n, which compose

[102] See Chalmers & Jackson (2001: 329-30).

the macroscopic entity *r*, then we can infer a priori the mass of *r*. This seems plausible because we are using the same predicate both at the microphysical and the macrophysical level, namely, 'mass'. But what happens when we introduce new predicates at higher-order levels? A complete description of the macrophysical level seems to require new predicates that did not appear at the microphysical level. Are we really able to apply these novel macrophysical concepts, given a microphysical description, without appealing to further empirical knowledge?

According to Chalmers and Jackson, we can indeed: "Overall, [PQ_T] implies complete information about the (geometrically characterized) structure and dynamics of macroscopic systems and objects in the world, their spatiotemporal distribution and microstructural composition, and their actual and potential perceptual appearances" (331). This macroscopic description seems to involve novel predicates that do not appear at a microphysical level, such as those describing the shape, size, behaviour and appearance of macrophysical objects. For illustration, let's consider the example 'square'. This predicate can be used to describe the geometrical properties of a macrophysical object. However, it is very unlikely that it appears at the microphysical level. So, how could we be in a position to find the extension of the concept SQUARE, given a description of the world in terms of PQ_T?

In order to be able to do that, the application conditional for the concept SQUARE should be something like this (as before, we will assume that the microscopic entities $x_1, x_2 \ldots x_n$ compose the macroscopic body *r*):

(AC_{square}) If $x_1, x_2 \ldots x_n$ instantiate properties $F_1, F_2 \ldots F_n$, then *r* is square.

(Here, properties $F_1, F_2 \ldots F_n$ are supposed to be microphysical properties). I think that application conditionals of this sort are required if we want to defend the view that PQ_T entails a priori a macrophysical description of the world. In particular, such a view requires the following assumption: in order to possess a concept from a certain level, we have to be able to apply the concept to a description in terms of *lower-level* concepts. Because, otherwise, how would we be able to infer macroscopic truths involving a concept at level *n*, from truths involving concepts at level *m* (where $m<n$)? If we want to bridge the gap between these two levels, we have to postulate at least some concepts at level *n* such that their application conditionals have antecedents that involve concepts from level *n*-1 (or below). Notice that this is

compatible with the fact that the application conditionals for some concepts at level *n* involve only other concepts from the same level; but my point is that, unless some of the concepts at level *n* have application conditionals involving lower-level concepts, we cannot infer truths about level *n* from lower-level truths.

The last step in the a priori entailment is the application of ordinary macroscopic concepts other than the macrophysical ones, given a description of the world in terms of macroscopic systems and their macrophysical properties:

> For example, knowledge of the appearance, behaviour, and composition of a certain body of matter in one's environment, along with complete knowledge of the appearance, behaviour, and composition of other bodies of matter in the environment, and knowledge of their relationship to oneself, puts one in a position to know (on rational reflection) whether the original system is a body of water. (Chalmers & Jackson (2001: 332))

Again, I think that it is necessary, for the second step of the a priori entailment to succeed, that our macroscopic concepts be associated with application conditionals involving lower-level concepts in the antecedent. Perhaps the a priori entailment is more complicated than this two-step entailment: maybe it involves more steps. However, it is crucial that at each step, our concepts can be applied a priori to descriptions in lower-level terms.

Therefore, the application conditional for these ordinary macroscopic concepts would be roughly as follows:

(AC_{macro}) If *r* has macrophysical properties H_1, H_2…H_n, then *r* falls under C.

These are the kind of application conditionals that Chalmers & Jackson require for their thesis to hold. Now we have to evaluate whether believing them is really necessary in order to possess the corresponding concepts, that is, whether it is true that concept possession requires, and therefore provides, the conditional abilities expressed by those application conditionals. For this, we have to consider whether there could be thinkers that were possessors of the relevant concepts and yet did not believe the corresponding application conditionals. Notice that we do not have to challenge Chalmers & Jackson's main assumption, namely, *ascriptivism* (i.e. the view that concept possession yields some kind of conditional ability to identify the extension of a concept, given a possible scenario). We can accept this and still

question whether we are actually able to identify the extension of our concepts a priori, given a scenario described in *lower-level terms*.

Therefore, we can see that Chalmers & Jackson are committed to something more than just the claim of ascriptivism. Following our considerations above, we can distinguish two different versions of ascriptivism: a weaker one, which I will call 'non-reductive ascriptivism', and a stronger one, which I will call 'reductive ascriptivism'. We can see that Chalmers & Jackson are committed to the stronger one, but as I will argue later, this stronger claim is very problematic.

The distinction amounts to this: In order to defend that $PQ_T \rightarrow W$ is a priori, Chalmers and Jackson have to add something to the ascriptivist thesis. They have to claim that application conditionals have the following crucial feature: for concepts at level *n*, feature F in the antecedent is described using concepts from a lower level *m*. I will call this view *reductive ascriptivism*:

> (**Reductive Ascriptivism**) For any concept C at level *n*, there is an application conditional like this: (AC): 'If *x* is F, then *x* falls under C', where feature F is described using concepts from a lower level *m*.

According to reductive ascriptivism, these application conditionals involving lower-level terms are part of the possession conditions of our concepts.[103] That is, in order to possess the concepts, we have to know those conditionals. Therefore, we can know those application conditionals *a priori*.

Then, a subject who possessed all the relevant concepts would know all the relevant application conditionals, and therefore, she would be able to *a priori* infer truths about any level from truths at the lowest-level, that is, the microphysical level.

However, we can accept the basic ideas of ascriptivism, and still deny the additional assumptions of reductive ascriptivism. For example, the following position is clearly ascriptivist but still falls short of being committed to reductive ascriptivism. I will call it *non-reductive* ascriptivism:

[103] As I explained before, it could be the case that for some concepts at level *n*, their application conditionals involve only other concepts at level *n*, but it is required that some of these other concepts have application conditionals whose antecedents do involve concepts at a lower level *m*, if we are going to be able to infer truths involving the former concepts from truths at level *m*. Therefore, I will focus on application conditionals whose antecedents involve lower-level concepts.

(**Non-reductive Ascriptivism**) For any concept C at level *n*, there is an application conditional like this: (AC): 'If *x* is F, then *x* falls under C', where feature F is described using only concepts at level *n*.

An advocate of non-reductive ascriptivism does not have to accept that microphysical truths a priori entail ordinary macroscopic truths such as truths about water. In particular, she does not have to accept that truths from level *n* are a priori entailed by truths from level *m*.

Let's consider again our example (W): 'Water covers 60% of the Earth's surface'. An advocate of non-reductive ascriptivism can coherently maintain that PQ_T&~W is conceivable. Hence, ascriptivism does not force us to maintain that $PQ_T \rightarrow W$ is a priori true.

In the following section, I will examine the main tenets of reductive ascriptivism, and I will show that it is wrong. Therefore, we can pose the argument against Chalmers & Jackson in the form of several dilemmas. First, we have to choose between ascriptivism and non-ascriptivism. If we choose the latter, then it is clear that $PQ_T \rightarrow W$ is not a priori true, although it is 1-necessary, so (CP) will turn out to be false. If, on the other hand, we choose ascriptivism, then we have to face another choice: either we choose reductive ascriptivism or non-reductive ascriptivism. Again, if we choose the latter, then $PQ_T \rightarrow W$ will not be a priori true, and then (CP) will turn out to be false as well. But if we choose reductive ascriptivism, then we will be committed to very implausible views about the possession conditions of our concepts, as I will show in the following section. Therefore, the best choice is to endorse non-reductive ascriptivism, if we want to endorse some form of ascriptivism. But then, $PQ_T \rightarrow W$ will pose a counterexample to (CP). Therefore, the prospects for the conceivability arguments are not good.

4.5. Problems for the A Priori Entailment Thesis: Concepts and Understanding

As we have seen, Chalmers and Jackson (2001) say that if a subject possesses a concept, then she is able to identify the concept's extension, given nontrivially sufficient information. This is the theory of concepts that they explain at the beginning of their paper, and the one we are assuming here. But as I have argued above, they are committed to something stronger: for certain macrophysical concepts,

microphysical information is *sufficient* information for finding out the extension of the concept, and for certain ordinary macroscopic concepts, macrophysical information is *sufficient* information in that sense too. I will argue that these claims are problematic because we can imagine hypothetical subjects who could be said to possess the relevant concepts and are nonetheless ignorant of the corresponding application conditionals of the sort required.

The application conditionals that are required for the a priori entailment thesis are roughly something like the following:

(AC_{square}) If x_1, x_2 … x_n instantiate properties F_1, F_2 … F_n, then *r* is square.
(AC_{macro}) If *r* has macrophysical properties H_1, H_2…H_n, then *r* falls under C.

In what follows, I will use these examples to illustrate the problem that reductive application conditionals face, in my view. What reasons could we have to assume that our concepts have application conditionals of this sort? For instance, let's focus on (AC_{square}). Do we have such an inferential ability, just in virtue of possessing the concept SQUARE? Can we apply such a concept to a lower-level description, without further empirical information?

Let's consider the case of a subject, Esther, who is not in a position to know the application conditional (AC_{square}). The crucial question is: does she possess the concept SQUARE? That is, is Esther able to entertain thoughts involving the concept SQUARE?

It seems perfectly compatible with lacking such conditional ability that she could in effect use the word 'square' very much as other subjects who have the conditional ability would. She can have conversations with those subjects, in which they apply the word 'square' to the same objects. If Esther did not possess the concept SQUARE, then how can she understand what the others are saying? How can she communicate with them? If we assume that linguistic communication takes place only if the speakers associate the same concepts with the same words, then the participants in those conversations have to express the same concepts with the same words in order to understand each other.[104] Then, if Esther can communicate with other more knowledgeable speakers, that is because they associate the term 'square' with the

[104] The principle (Have) below points in this direction.

same concept, and therefore, she has the concept SQUARE even if she does not have the ability to apply it to a microphysical description. Therefore, this ability is not part of the concept-possession conditions, and therefore, (AC_{square}) is not a priori.

We can put the problem in this way: if we assumed that in order to possess macroscopic concepts such as SQUARE we have to associate them with reductive application conditionals such as (AC_{square}), then it would follow that we lack many concepts that we thought we did possess. We all seem to possess macrophysical concepts such as SQUARE, ROUND, LENGTH, PRESSURE and so on. (At least, we can use the corresponding words competently, and this strongly suggests we do possess such concepts.) However, if in order to possess them we had to be able to infer truths involving them from lower-level descriptions, it would turn out that we do not really possess such ordinary concepts, since we do not seem to have such sophisticated inferential capacities.

The same point applies to the rest of the reductive application conditionals that are required by Chalmers and Jackson's view. We can easily think of subjects who do not have the relevant conditional abilities that are required by the a priori entailment thesis, and at the same time, these subjects seem to understand the relevant words. This does not have to conflict with our general assumption that understanding of words requires conditional abilities; it is just that those subjects have less sophisticated conditional abilities. For instance, it is not clear at all that we have to know a priori application conditionals such as (AC_{macro}) in order to understand the corresponding words for ordinary macroscopic concepts. It is quite usual for a subject to be ignorant of the macrophysical properties that determine the instantiation of ordinary macroscopic properties, but we still attribute linguistic understanding of the corresponding words to these subjects.

My point can be put like this. If the following principle (Have)—proposed by Timothy Williamson (2003)—is assumed, then the view about concepts we are evaluating will be in jeopardy:

> (Have) If one understands the word 'C', one has the concept C. (Williamson (2003: 290))

This principle seems plausible enough. But if we accept it, then there is a problem for the view that the application conditionals above are part of the possession conditions

for the relevant concepts. The problem is that it seems that there are subjects who are ignorant of those conditionals but are able to understand the corresponding words. By (Have), they therefore possess the corresponding concepts. Therefore, those conditionals are not part of the concepts' possession conditions. And therefore, we cannot infer a priori M from PQ_T.

There is a complication that we should examine, though. Chalmers and Jackson explain that they are focusing on expert subjects, that is, subjects that have a full, nondeferential grasp of their concepts,[105] and this could bring out problems for the line of response I am putting forward. Let me elaborate on this.

What does it mean to say that a certain subject has a deferential or a non-deferential grasp of a concept? When we use concepts deferentially, we defer to the experts' abilities to identify the extension of the concept. Then, our own conditional abilities consist in being able to identify the relevant extension, given a description of a scenario in terms of how the experts use such a concept. For instance, we use WATER deferentially when we do not know what properties determine that something falls under WATER (in a possible world considered as actual) but we do know at least that there are experts who know those properties. Hence, we could identify water in a given scenario provided that we knew what the experts say about the extension of water. On the other hand, non-deferential users of a word (that is, *experts*) are those who know what properties determine whether something falls under a certain concept, and therefore they could find out the extension of such a concept, given a sufficiently rich description of a scenario.

Therefore, we could distinguish between *full* grasp and *deferential* grasp of the meaning of a word (or the concept expressed by the word). According to this view, full grasp would require a conditional ability to apply the concept to a lower-level description of the world (at least, for some concepts), whereas deferential grasp would only require deference to experts' abilities. Then, Chalmers and Jackson can reformulate their a priori entailment thesis as claiming that for all thinkers with *full grasp* of their concepts, PQ_T entails a priori M.

Hence, this strategy is committed to the claim that in order to fully grasp the meaning of our words (and possess the corresponding concepts), we have to know the application conditionals of the sort postulated. The subjects who know these

[105] See Chalmers & Jackson (2001: 328).

application conditionals will have a full grasp of the concepts. Subjects who are ignorant of these application conditionals can still have deferential grasp of the concepts, if they rely appropriately on the experts.

One initial problem with this view is the following: what if there are no experts on a certain subject matter? How could we distinguish between deferential and full grasp of the relevant concepts then? We have said that deferential grasp of a concept relies on the experts' use of the concept. But then, if there are no experts on a certain matter, it follows that no-one fully possesses the corresponding concepts, and furthermore, that no-one can even have a partial grasp of a concept, since there are no experts to rely on.

However, we can put this problem aside, since in my view, there are two further and more serious problems with Chalmers and Jackson's view (according to which only experts have full mastery of the concept/word), which I will now explain in turn.

The first of these problems is that if we endorse Chalmers & Jackson's view, then it would follow that almost any piece of empirical knowledge can become a priori knowledge. To illustrate this problem, let's examine some examples of reductive application conditionals that, according to that view, should be known in order to possess the corresponding concepts. Let's consider, for instance, the macroscopic concept CONDENSATION (i.e. the change of vapour into a liquid). We can assume that the relevant application conditional for such concept would be something like this:

> ($AC_{condensation}$): If x_1, x_2 … x_n instantiate properties F_1, F_2 … F_n, then *r* condenses.

Let's now consider a subject, Gab, who is at the beginning of a degree in chemistry. During her first weeks at college, she does not know yet what microphysical properties determine the condensation of vapour. That is, she does not know ($AC_{condensation}$).

Therefore, according to the view that we are considering, Gab does not fully possess the concept CONDENSATION, nor does she fully grasp the meaning of the expression 'condensation': she has to learn more facts in order to fully grasp the concept and understand the term. Gab needs to acquire more information in order to

fully understand the expression. Then, the following question arises: at what point does she become able to fully grasp such expression?

The problem is the following: according to this view, Gab has to learn some empirical facts about condensation, for instance by attending lectures on chemistry, so as to fully grasp the meaning of the expression 'condensation'.[106] However, according to Chalmers and Jackson, once she has learned these empirical facts about condensation, this knowledge will become part of the possession conditions of the concept CONDENSATION (or at least, part of the conditions for *fully* possessing the concept), and therefore, that knowledge will become *a priori* knowledge. This is the aspect of this position that I find problematic: it is held that certain knowledge is *a priori* for a subject, but at the same time we have seen that this subject has to acquire that knowledge in a way that resembles paradigmatic *a posteriori* knowledge, such as going to chemistry lectures. The problem is that, if the a priori entailment thesis is correct, then this knowledge becomes *a priori* knowledge for Gab. That is, she has to know the relevant application conditional ($AC_{condensation}$) a priori, because otherwise she could not a priori infer truths about the condensation of macroscopic substance *r* from PQ_T. But, how could such knowledge suddenly become *a priori* knowledge? Going to chemistry lectures seems to be a paradigm of coming to know something a posteriori. The point here is not only that we typically come to know that sort of information in an a posteriori manner. It could be argued that we can come to know mathematical truths such as '7+5=12' in an a posteriori manner too (say, by being taught that in school), but that these truths can also be known a priori. The problem

[106] It could be argued that what is relevant here is whether the information encapsulated in the *primary intension* of a term such as 'condensation' is known a priori or not. Therefore, perhaps the empirical facts about condensation that we are referring to are not relevant here: what is relevant is the causal-functional role that we associate with 'condensation', and not the microscopic properties that happen to realize that role in the actual world. The main idea here is that the information that is relevant in order to know the application conditional for CONDENSATION (and therefore being an expert concerning such a concept) is information about the causal-functional role of condensation, not about the microscopic properties responsible for condensation at a chemical level. My reply to this objection is twofold. On the one hand, I think that Chalmers & Jackson's thesis of a priori entailment requires that we are able to apply some macrophysical concepts to a microphysical description, and in order to do that it is not enough that we know the causal-functional roles associated with those macrophysical concepts: we also have to be able to apply these roles to a microphysical description, and it seems that we need to be able to apply some macrophysical concepts to a microphysical description, in order to do that. On the other hand, we are assuming here that for many physical concepts (those invoked in P) the primary and the secondary intensions coincide, and therefore knowing the primary intension would require knowing the secondary intension. In any case, it seems clear that Chalmers & Jackson are committed to the view that in order to be experts concerning many macrophysical concepts, we have to be able to apply them to a microphysical description, and this seems to require the kind of application conditionals that we are discussing here.

here is that there seems to be no way that we could come to know the information that is relevant in order to know the corresponding application conditionals in an a priori manner. For how could we come to know that such and such microphysical properties determine a body's condensation, merely a priori?

The basic problem with the idea that such application conditionals are a priori is that this would amount to *trivialising* the notion of a priori knowledge. We can illustrate this with another example. Imagine we discover that certain property K is instantiated by something if and only if it instantiates properties X, Y and Z. Imagine also that we have a concept C* to refer to K. It would seem that the following sentence would express an interesting discovery:

(AC_{C*}) If *x* has X, Y and Z, then *x* falls under C*.

However, what happens when we apply Chalmers and Jackson's view to this case? What would be required for a subject to have full grasp of such a concept C*? Well, they are committed to the claim that in order to fully grasp a concept, we have to be able to find its extension in a description in lower-level terms, and therefore this means that we would have to know something like (AC_{C*}) a priori. That is, this is the sort of reductive application conditional that, according to them, we have to know in order to fully possess the corresponding concept. But then, it follows that all empirical discoveries of this sort would become a priori knowledge. This seems to be an undesirable result: is a priori knowledge really that easy to get?

But again, we can put this problem aside, because there is third problem for Chalmers and Jackson's view that is even more serious. The problem is not just that the notion of apriority that they use is such that a priori knowledge would become much more common than we usually assume it is. The more worrying problem is that if we take into account such notion of apriority, then we have no grounds to think that the conditional 'If P, then Q' is not a priori. Let me explain this point a bit more slowly.

We can reconstruct Chalmers and Jackson's argument with respect to the previous example as follows: First, (i) they would claim that what makes someone an *expert* concerning C* is that they are in a position to know (AC_{C*}). From this, Chalmers and Jackson infer (ii) that knowing (AC_{C*}) is part of the possession conditions of concept C*, or at least, that in order to *fully possess* the concept, we

have to know (AC_{C*}). From this, they infer that (iii) such application conditional can be known *a priori*.

The crucial point here is that this notion of a priori does not seem to be the substantial notion of a priori knowledge that is at play in discussions of conceivability arguments. This substantial notion has it that a priori knowledge is that knowledge which can be justified independently of experience. However, we can see that the notion of a priori knowledge involved in (iii) is a more trivial notion. We can always stipulate that in order to possess concepts to a certain degree, we have to know certain empirical information. Then it trivially follows that we know that information just *by virtue* of possessing the concepts to a certain degree. In this sense, that knowledge is a priori: once you possess the concepts, you do not need *further* experience in order to know such information. But this clearly does not mean that the relevant piece of knowledge is a priori in the standard sense of the term. In particular, as we have seen in this case, the application conditional (AC_{C*}) has been discovered by empirical means.

Therefore, it seems that the notion of apriority that Chalmers and Jackson are invoking in their argument (at least, if they want the argument to be valid, that is, if they want (iiii) to follow from (i) and (ii)) is a *technical* notion of apriority, according to which *p* is a priori when *p* is that piece of information that subjects need to know in order to be *experts* concerning C*. As we have seen, in order to be an expert concerning C*, according to Chalmers and Jackson, it is not enough that we use C* competently in conversations and so on, but we have to be in a position to know certain empirical information about what lower-level properties of the referent of C* determine that it falls under C*.

As I have said earlier, I think that such empirical knowledge cannot constitute a priori knowledge, precisely because it cannot be known independently of experience, and if we assumed that it is a priori knowledge, then this would trivialize the notion of the a priori. But this claim is not really essential in order to see that Chalmers & Jackson's strategy cannot work. What is crucial is that their notion of apriority cannot do the work that they intend it to do. Let me review the dialectic here in order to show why this is the case.

Chalmers and Jackson's aim is to argue that (a) microphysical truths (plus phenomenal truths) entail ordinary macroscopic truths *a priori*, and then use this claim to argue that (b) if phenomenal truths are not *a priori* entailed by microphysics,

then phenomenal truths cannot globally supervene on the physical. The problem with this argument can be put like this: the only way of making (a) remotely plausible is by appealing to their technical notion of a priority, but this is not the notion appealed to in (b), and therefore, they cannot use (a) to support (b).

In particular, if we understand (b) according to their technical notion of a priori, then the antecedent of (b), that is, the claim that microphysical truths (P) do not entail phenomenal truths (Q) a priori, would no longer be plausible. This claim was plausible and intuitively compelling because we were understanding 'a priori' in the standard sense, that is, as knowledge that can be justified independently of experience. And in this sense, it seems clear that physical truths do not entail phenomenal truths a priori. But if we invoke Chalmers and Jackson's trivialised notion of a priori knowledge, then our case for the claim that P does not entail Q a priori would lose its intuitive force. Our grounds for that claim have to do with our intuitions concerning the conceivability of zombies, zombie worlds, and so on, but these intuitions are completely irrelevant to Chalmers and Jackson's technical notion of apriority. In particular, as we have seen, in order to be in a position to find out whether P entails Q a priori *in that sense*, we would have to be *experts* concerning the concepts involved in Q, and in order to be experts in Chalmers and Jackson's sense, we would have to know certain sophisticated application conditionals of a *reductive* sort. And unfortunately, many of us are not in a position to know such things. Furthermore, it is controversial whether such reductive application conditionals could be available at all (since, according to some people, a reductive explanation of phenomenal properties is just not possible).

Therefore, coming back to the main point, if we wanted to use (a) above to support (b), then we would have to appeal to the same notion of apriority in both claims, but if we use the standard notion of apriority, then (a) is not plausible, and if we use Chalmers and Jackson's technical notion, perhaps (a) would be plausible but then we would no longer have any grounds for claiming that physical truths do not entail phenomenal truths a priori. And therefore, the conceivability argument against physicalism would collapse.

Therefore, the notion of a priority that is relevant with respect to (b) is the standard notion of knowledge that can be justified independently of experience. However, as I have argued, we have good reasons to believe that we cannot infer macroscopic truths such as W from PQ_T a priori in that sense, because in order to

make this inference we would need to have certain pieces of *empirical* information that only experts know. Therefore, we can conclude that Chalmers and Jackson's claim that PQ_T entails ordinary macroscopic truths M *a priori* is wrong.

4.6. Conclusion

Recapitulating, I have argued that Chalmers and Jackson's assumptions about concepts (in a nutshell, that possessing macroscopic concepts enables the thinker to infer macroscopic truths from a description of the world in the language of physics) are very problematic. I have proposed some counterexamples in which a subject lacks those inferential abilities but she intuitively could be said to understand the words for those concepts and therefore possess the concepts. These examples show that such inferential abilities are not really part of the possession conditions of our concepts. And this entails that we are not really able to infer ordinary macroscopic truths such as W from a description of the world in terms of PQ_T *a priori* (in the relevant sense). As we have seen, these inferences could be a priori only if our concepts were a priori associated with application conditionals of a reductive sort. Since they do not seem to be so associated, then PQ_T cannot entail W a priori.

Therefore, we can assert that the conditional $PQ_T \rightarrow W$ is not a priori true. But we have reasons to assert that it is true in all possible worlds considered as actual. Therefore, we can conclude that it is a clear case of strong necessity. Hence, its negation $PQ_T \& \sim W$ is a conceivable impossibility, and this entails that the conceivability-possibility principle (CP) is incorrect.

Let's take stock. In this chapter, we have focused on the two-dimensional argument against physicalism. But in Ch. 3 we introduced a related argument, namely, the two-dimensional argument against identity theses. Let's recall it here for convenience:

The Two-Dimensional Argument against Type-identities

I. The identity-statement 'pain is C-fibre firing' is a posteriori.
II. If an identity statement 'A=B' is a posteriori, then the terms 'A' and 'B' have different primary intensions. (DPIM)
III. Therefore, 'pain' and 'C-fibre firing' have different primary intensions.
IV. The primary and secondary intensions of 'pain' coincide.

V. The primary and secondary intensions of ‘C-fibre firing’ coincide.

VI. Therefore, ‘pain’ and ‘C-fibre firing’ have different secondary intensions.

VII. If ‘pain’ and ‘C-fibre firing’ have different secondary intensions, then ‘pain is C-fibre firing’ is false.

———————

VIII. ‘Pain is C-fibre firing’ is false.

As we saw in Ch. 2, the crucial premise (CP) of the two-dimensional argument against physicalism is entailed by the core thesis of two-dimensional semantics, namely, (2D):

> (2D) A sentence is a priori if and only if it is true in all possible worlds considered as actual.

In this chapter we have explored the so-called non-exceptionalist strategy against (CP), and we have concluded that there are good reasons to think that there are counterexamples to (CP). We can conclude, then, that (2D) is false too, since (2D) entails (CP).

What about the *Distinct Primary Intension Model* (DPIM), which is crucial for the argument against type-identities above? As we saw in 3.1, this principle is also supported by (2D). Actually, (DPIM) is just an application of (2D) to the case of identity statements. According to (2D), a sentence is a priori if and only if it is true in every possible world considered as actual. This means that an identity statement will be a priori if and only if both terms of the identity statement refer to the same entity in all possible worlds considered as actual—that is, when both terms have the same primary intensions. And this entails that if an identity statement is a posteriori, the corresponding terms have different primary intensions, which is exactly what (DPIM) says.

If the non-exceptionalist strategy that we have examined here is correct, then it means that (2D) is wrong, and therefore (DPIM) now lacks support. For if we can provide examples of sentences that are a posteriori but true at every possible world considered as actual (such as $PQ_T \rightarrow M$), then there is no reason at all to move from the identity statement ‘Pain is C-fibre firing’ being a posteriori to its being false at some

possible world considered as actual. Therefore, the proper conclusion of the non-exceptionalist strategy, if successful, is that the two-dimensional argument against identity theses above is not correct either.

5. The Exceptionalist Strategy I: Introducing the Recognitional Account

5.1. Introduction

In this chapter, we will explore the so-called exceptionalist strategy against the two-dimensional arguments. Let's recall both arguments here again for convenience:

The Two-Dimensional Argument against Physicalism

2DA1: P&~Q is conceivable.

(CP): If sentence S is (ideally) conceivable, then S is primarily possible.

2DA2: If P&~Q is conceivable, P&~Q is primarily possible.

2DA3: If P&~Q is primarily possible, P&~Q is secondarily possible.

2DA4: If P&~Q is secondarily possible, Physicalism is false.

_____

2DA5: Physicalism is false

The Two-Dimensional Argument against Type-identities

I. The identity-statement 'pain is C-fibre firing' is a posteriori.
II. If an identity statement A=B is a posteriori, then the terms A and B have different primary intensions. (DPIM)
III. Therefore, 'pain' and 'C-fibre firing' have different primary intensions.
IV. The primary and secondary intensions of 'pain' coincide.
V. The primary and secondary intensions of 'C-fibre firing' coincide.
VI. Therefore, 'pain' and 'C-fibre firing' have different secondary intensions.
VII. If 'pain' and 'C-fibre firing' have different secondary intensions, then 'pain is C-fibre firing' is false.

_____

VIII. 'Pain is C-fibre firing' is false.

We saw in the previous chapter that the non-exceptionalist strategy can provide a very successful response against both sorts of arguments. Then, why should we examine yet another line of response? I think that we have good reasons to examine the

exceptionalist strategy, since in this way we can see that conceivability arguments are wrong in both counts.

In the previous chapters I have explored the prospects for conceivability arguments, given different background views about concepts and concept-possession: for instance, we have explored views as diverse as analytic functionalism, non-ascriptivism and non-reductive ascriptivism. We have seen that on any of these views, it can be shown that conceivability arguments are unsuccessful. The only view that supports conceivability arguments, namely, what I called 'reductive ascriptivism', was shown to be problematic in the previous chapter. I believe that those criticisms are in the right track, and that the problems for reductive ascriptivism pose a big challenge to the advocate of conceivability arguments. However, it is still in my view worth exploring the prospects of the exceptionalist strategy against conceivability arguments, since, crucially, this is a strategy which shows that even if reductive ascriptivism was ultimately correct conceivability arguments would still fail.

In this way, the overall strategy against conceivability arguments that would ensue from our discussion would be extremely powerful: we would have surveyed many different views that are available concerning theories of concepts (both phenomenal concepts and the rest of empirical concepts), and we could conclude that, on any of these views, it follows that conceivability arguments fail. This is a very important result, which has been overlooked in previous discussions of conceivability arguments, which tend to focus on some particular view on concepts.[107]

Therefore, in what follows I will discuss the exceptionalist strategy in large detail. I will start by introducing its main tenets, and I will then go on to introduce one of the most prominent versions, namely, the so-called *recognitional* account. This version has been developed by Brian Loar, and Chris Hill & Brian McLaughlin, among others.[108] After that, I will present some recent objections to the recognitional account by Daniel Stoljar, and I will argue that they do not work. In the next chapter, I will examine some more general objections to the exceptionalist strategy under any of its guises, first by Stoljar, and then by Chalmers. I will argue that neither of these objections really works. Along the way, I will clarify the aims and structure of the exceptionalist strategy, and I will develop a version which, in my view, can satisfy

[107] One important exception is Levine (2001), which, although focusing on non-ascriptivism, also explores, even if briefly, the prospects of the ascriptivist view with respect to conceivability arguments.
[108] See Loar (1990), (1997), (1999) and (2003). See also Hill & McLaughlin (1999).

these aims. Hence, I will conclude, the exceptionalist strategy is a successful strategy against conceivability arguments, and these irredeemably fail.

5.2. The Exceptionalist Strategy: Phenomenal Concepts and Conceivability Arguments

There is a certain degree of agreement between many exceptionalists and the proponents of conceivability arguments. The thesis they agree upon is *ascriptivism*, in the sense we gave to the term in the previous chapter. [109] That is, they think that most concepts are associated a priori with modes of presentation that enable subjects to identify their extensions in different scenarios.

For instance, Loar is an *ascriptivist* in this sense. In particular, he is a *reductive* ascriptivist (concerning non-phenomenal macroscopic concepts), so he thinks it is possible to infer non-phenomenal macroscopic truths from microphysical truths (suitably qualified) a priori. For instance, he says that "when we explain liquidity […] in physical-functional terms, the explanation is in crucial part *a priori*. […] What we in effect do is to analyse liquidity […] in terms of a functional description, and then show that the physical theory of liquids implies, *a priori*, that the functional description is realized" (Loar (1999: 470)).

Then, according to Loar and other exceptionalists, sentences of the form P→M (where M is a macroscopic truth, such as a truth about liquidity) are a priori.[110] So they cannot be strong *a posteriori* necessities. In other words, P&~M is not conceivable. Therefore, they do not accept that sentences of this sort constitute counterexamples to (CP).

The exceptionalist response against (CP) and the two-dimensional argument is different. Their advocates argue that (CP) fails to hold when it comes to sentences that involve *phenomenal concepts*; likewise with respect to (DPIM). This is the core idea of the exceptionalist strategy.

---

[109] Some prominent advocates of exceptionalism, such as Loar, clearly agree with the proponents of the conceivability argument in this respect, but it is not clear that they all do.

[110] According to Chalmers and Jackson, as we saw in the previous chapter, the conditionals that are clearly a priori are of the form $PQ_T \rightarrow M$ (where P includes T (the "that's all" clause) and I (the indexical information), and Q_T is a complete phenomenal description of the world). For it is controversial whether $P \rightarrow Q_T$ is a priori, and there are many truths M that cannot be inferred from P unless we add Q_T. We will ignore this complication in what follows.

Exceptionalists accept that P&~Q is conceivable, that is, it is not a priori false. They also accept that identity statements between phenomenal concepts and physical concepts, such as (3), are typically a posteriori:

(3): Pain is C-fibre firing

The question is whether these epistemic features pose a problem for physicalism or for such identity-theses. Exceptionalists think they do not.

In particular, exceptionalists want to defend the coherence of *type-B materialism*, and therefore they want to defend the coherence of the view that P&~Q is conceivable but false at all possible worlds considered as actual, or in other words, that P→Q is a posteriori but true in all possible worlds considered as actual (1-necessary). Likewise, type-B materialists claim that identity statements such as (3) are a posteriori but true. What the exceptionalist tries to do is to defend the *coherence* of these views (therefore, she does not have to offer positive arguments for the truth of these views). [111] The strategy works as follows.

As we have seen, the proponents of the two-dimensional arguments rely on certain principles, that is, (CP) and (DPIM) (both of which depend on (2D)). These models have been developed independently of the question of physicalism, but they have consequences that are relevant for the question of physicalism (and identity theses). In particular, (CP) entails that, if the sentence P&~Q was conceivable, then it would be true in some possible world considered as actual (1-possible). Concerning psychophysical identities such as (3), (DPIM) entails that if they are a posteriori, the corresponding terms will be associated with different primary intensions. The principles (CP) and (DPIM) in themselves are neutral concerning physicalism and the truth of psychophysical identities such as (3); they just predict what would happen if those theses were true, and what would happen if they were false. If physicalism was true, then P&~Q would not be 1-possible, and therefore, according to (CP), it would not be conceivable either. In addition, according to (DPIM), if the identity statement

[111] Barbara Montero (2003) has offered a criticism of exceptionalism, according to which the exceptionalist strategy does not succeed because it fails in providing evidence for the claim that physicalism is true. In response, I think that exceptionalism does not have to provide evidence for the claim that physicalism is true. The proper aim of the strategy is to show that for all we know, including the fact that (3) is a posteriori, phenomenal properties *coul*d be identical to physical properties. That is, exceptionalists have to show that (3) being a posteriori is compatible with it being necessary; they are not required to show that (3) is necessary.

(3) was true then it should be a priori true. [112] Given (2D), and the additional claim that P&~Q is conceivable, it follows that P→Q is not 1-necessary, and therefore (given certain assumptions), physicalism is false. Likewise, it follows from (2D) plus the claim that (3) is a posteriori that (3) is not 1-necessary and therefore is false.

The exceptionalist's task is to propose an *alternative* framework to deal with phenomenal concepts, such that it predicts what would happen if physicalism (and also psychophysical identity statements) were true. (This framework should be justified independently of the question of whether such claims are indeed true.) And then we can apply it to the relevant sentences, to see what the consequences are.

Most versions of the exceptionalist strategy put forward arguments with the following structure:

a) A certain theory of phenomenal concepts C is true of human beings.
b) If phenomenal concepts satisfy account C, then even if P&~Q were 1-impossible (that is, false at all possible worlds considered as actual), it would still be conceivable.
c) So the conceivability of P&~Q does not show that it is 1-possible.

The conclusion of this argument is that the conceivability of P&~Q does not entail its 1-possibility, and therefore (CP) is wrong. Notice that this strategy does not proceed by offering counterexamples to (CP). The strategy is not committed to P&~Q being a counterexample to (CP). What the strategy asserts is that a) a certain account of phenomenal concepts C is correct, and b) it follows from C that, if P&~Q was 1-impossible, it would still be conceivable. This is neutral concerning whether P&~Q is in effect 1-possible or not. The upshot of this strategy is to show that the inference from a sentence's conceivability to its 1-possibility is unjustified.

Likewise, we could also argue against (DPIM) as follows:

a) A certain theory of phenomenal concepts C is true of human beings.

[112] Assuming, of course, that 'pain' has the same primary and secondary intensions, and likewise for 'C-fibre firing'. If this assumption is dropped, then it would be consistent with (DPIM) to say that (3) is both a posteriori and true.

d) If phenomenal concepts satisfy account C, then identity statements such as (3) would not be a priori, even if they were 1-necessary (i.e. true at all possible worlds considered as actual).
e) So the aposteriority of (3) does not show that (3) is not 1-necessary.

According to this presentation of the strategy, the crucial tasks for the exceptionalist are, on the one hand, to find an account of phenomenal concepts that satisfies b) and d), and on the other, to motivate the view that this account is *the correct one*, as a) says.

However, it is important to notice that the exceptionalist strategy can still be very useful even if it does not accomplish the second task, that is, even if it does not show that the corresponding account of phenomenal concepts in fact applies to human beings. This is so because principles (CP) and (DPIM) are submitted as *a priori principles*. That is, their advocates argue that they are a priori true and that the arguments offered in support of them are *a priori arguments*. For instance, Chalmers (1999) is explicit on this point: "I hold that CPT [the conceivability-possibility thesis] is *a priori*, although highly nontrivial, like many theses in philosophy" (1999: 485). If so, then we just need to show that b) above is correct, in order to show that (CP) is not a priori true. That is, if there is a *coherent* account of phenomenal concepts which entails that even if P→Q is true at all possible worlds considered as actual, it will not be a priori true, then this means that there is a *conceivable* situation under which conceivability does not entail 1-possibility, and this would entail that (CP) is not a priori true, *contra* Chalmers. And *mutatis mutandis*, if we show that d) above is correct on some account of phenomenal concepts C, then it follows that (DPIM) is not a priori true. And if we show that such principles are not a priori true, then it is not possible to build up an argument against physicalism (or identity theses) on merely *conceptual* grounds. The conceivability arguments that we have been considering aim to show that physicalism is false on conceptual grounds. That is, the justification for the principles (CP) and (DPIM) is supposed to be a priori, as well as the rest of the argument. The only premise of the conceivability argument that is not intended to be a priori is the claim that there are phenomenal truths in the actual world. This can be an a posteriori truth, even if, arguably, it can be known by introspection. Then, the whole conceivability argument could be rehearsed "from the armchair", so to speak. This conception of the conceivability argument will be clearly damaged if we show that the

corresponding principles (CP) and (DPIM) are not a priori true. This would show that it is not possible to falsify physicalism (nor identity theses) by means of *conceptual methods*.

In any case, my aim is not only to show that (CP) and (DPIM) are not true a priori, but also and more crucially that they are false. And in order to do this, we need to find an account of phenomenal concepts which is true of human beings, and in terms of which we can defend b) and d) above. In the next section, I will present the *recognitional* account of phenomenal concepts, which in my view is a very plausible account of the nature of our phenomenal concepts. I will explain what features are attributed to our phenomenal concepts, according to this account, and why I think that our phenomenal concepts do possess these features. In addition, I will explain how we can defend claims b) and d) above in terms of this recognitional account of phenomenal concepts.[113]

5.3. In Search of Alternative Explanations: Loar's Recognitional Account

In this section I will introduce Loar's version of the exceptionalist strategy in terms of his recognitional account of phenomenal concepts, which clearly falls under the model sketched above. In general terms, Loar argues that, if we assume that phenomenal concepts refer to physical properties, then the most plausible theory about phenomenal concepts would entail that sentences such as (3) are still a posteriori. He focuses on the view that phenomenal properties are identical to certain physical properties of the brain, so that phenomenal concepts co-refer with some

[113] Other accounts of phenomenal concepts that have been suggested by advocates of the exceptionalist strategy are: the *quotational* account (Papineau (2002) and (2007)), the *indexical* account (Perry (2001)), and the *epistemic* or *self-presenting* account (Sturgeon (2000), Hill & McLaughlin (1999)). I will briefly discuss some of these in what follows, but detailed discussion of them is beyond our scope here, since my aim is to defend the recognitional account. Notice that this account is prima facie compatible with these other accounts, on two counts: first, the claim that phenomenal concepts have a certain feature R posited by a certain account is compatible with the claim that they also have another feature R' posited by some other account. So there is no obvious reason to think that if (some of) these other accounts mentioned above turned out to be correct, the recognitional account would not be correct. Secondly, the claim that we can defend b) and d) in terms of the recognitional account is also prima facie compatible with the claim that we can defend b) and d) in terms of some other accounts. For instance, it could be argued that phenomenal concepts having feature R entails that P&~Q would be conceivable even if it was impossible, but this is perfectly compatible with the claim that phenomenal concepts having the different feature R' also entails that P&~Q would be conceivable even if it was impossible. Therefore, my defence of the exceptionalist strategy in terms of the recognitional account is prima facie compatible with other accounts of phenomenal concepts, and with other versions of the exceptionalist strategy. I hope to explore further the prospects of some of these other versions elsewhere.

physical concepts. He argues that the type-identity view does not entail that a phenomenal concept and its corresponding coreferential physical concept should be a priori connected. If this is so, then a sentence such as (3) could be both true and a posteriori. And also, its negation could be conceivably true but 1-impossible. So b) and d) above would be vindicated.

Let me emphasise that this argument does not presuppose that phenomenal concepts refer to physical properties in a question-begging manner. Conceivability arguments aim to show that if phenomenal concepts referred to physical properties, certain epistemic consequences would occur (for instance, (3) would be a priori). Since such consequences do not seem to obtain, they conclude that phenomenal concepts do not refer to physical properties. Loar replies that, on the contrary, if phenomenal concepts referred to physical properties, those alleged epistemic consequences would not really follow, so their failure to take place does not pose a problem against physicalism.

An important feature of his response is that he focuses on *identity claims*, such as (3). His physicalist position has it that phenomenal properties are identical to physical properties, instead of being merely globally supervenient on physical properties, as the standard physicalist position would claim. Hence, Loar focuses mainly on the two-dimensional argument against type-identity theses, although what he says can easily be applied to the argument against physicalism.

The central question, as we have seen, is whether there is any reasonable account of phenomenal concepts in terms of which we can offer an alternative explanation of the epistemic gap between physical truths and phenomenal truths (that is, that P does not entail Q a priori, or that psychophysical identities are always a posteriori). The exceptionalist strategy claims that the epistemic gap is not due to any ontological gap, as advocates of the conceivability argument suggest, but rather to the special features of phenomenal concepts. For this reason, the exceptionalist strategy has also been labelled *the phenomenal concept strategy*.[114]

The exceptionalist strategy claims that it is because phenomenal concepts have certain key features that physical concepts and phenomenal concepts are not *a priori* connected. This lack of a priori connection means that we can conceive of a given

[114] This label is introduced in Macdonald (2004), although here it is used just to refer to some particular versions of the strategy. Stoljar (2005b) uses the term to refer to all of the different responses to conceivability arguments that attempt to explain the epistemic gap between P and Q in terms of the special features of phenomenal concepts, and I will be using the label in this way too.

phenomenal concept applying to something without any physical concept applying to it, and *vice versa*. That is, thoughts that consist in the joint application of a phenomenal and physical concept to an individual are always *a posteriori*.

The main task, then, is to explain the a posteriori status of sentences such as (3) without invoking different primary intensions associated with the terms, but just appealing to the special nature of phenomenal concepts. Loar has offered a promising explanation of the aposteriority of (3) along these lines.

Loar's main idea concerning the character of phenomenal concepts is that they are *recognitional concepts*. Loar introduces the idea of a recognitional concept by means of the following example: "Suppose you go into the California desert and spot a succulent never seen before. You become adept at recognizing instances, and gain a recognitional command of their kind, without a name for it; you are disposed to identify positive and negative instances and thereby pick out a kind" (Loar (1997: 600)). In this case, a recognitional concept is formed: the subject has the ability to have thoughts about that kind of succulent, by virtue of having the ability to recognise instances of it. Recognitional concepts, Loar says, "are grounded in dispositions to classify, by way of perceptual discriminations, certain objects, events, situations" (*Ibid*).

So, what about phenomenal concepts? Is it plausible to say that phenomenal concepts are recognitional concepts? In my view, we have good reasons to think so. First, the recognitional account can easily explain the following feature of our phenomenal concepts: in order to possess a phenomenal concept, we typically need to have had the corresponding experience before. For instance, in order to have the phenomenal concept 'experience of red', we need to have had experiences of red. This fact can be easily explained if we assume that phenomenal concepts are recognitional concepts (that is, associated with dispositional abilities to recognise new instances of the corresponding experience). In particular, the explanation would go like this: it is usually the case that in order to get the ability to recognise instances of a certain kind of experience, we have to have such experience. Then, if phenomenal concepts are just these recognitional abilities, it is clear why in order to get the phenomenal concepts, we typically need to have had the corresponding experiences, since we typically need to have had an experience of the corresponding kind, in order to get such recognitional abilities.

However, it is important to notice that having had an experience of a certain kind is not always necessary in order to form a phenomenal concept *of that kind*. For we can sometimes acquire *complex* phenomenal concepts just by virtue of possessing their constituent phenomenal concepts. For instance, if I possess the phenomenal concepts 'experience of seeing red' and 'experience of seeing a sphere', I could easily thereby acquire the concept 'experience of seeing a red sphere'. Again, this has a simple explanation on the recognitional account, which could go like this: we can acquire the complex concept just by virtue of having the constituent concepts, because having the constituent concepts (that is, the ability to recognise instances of seeing red and instances of seeing spheres, respectively) would enable us to *recognise* instances of the experience of seeing a red sphere. And this is exactly what is required in order to possess the complex concept of seeing red spheres.

Another good reason to think that phenomenal concepts are recognitional concepts is that this provides us with a very good explanation of how phenomenal concepts can get their referents fixed. Let me elaborate this point a little.

The main idea of the recognitional account is that every concept of phenomenal type K is associated with a disposition to re-identify tokens of phenomenal type K. For instance, the phenomenal concept PAIN is associated with a disposition to discriminate tokens of pain. Phenomenal concepts of more specific kinds of pain are associated with dispositions to re-identify these more specific types in pain-tokens. For instance, the concept HEADACHE is associated with the dispositional ability to discriminate tokens of headache from other tokens of pain.

What grounds these dispositional abilities to discriminate phenomenal tokens as being of the same kind? Loar suggests that phenomenal concepts incorporate phenomenal tokens, or sometimes just a memory of a phenomenal token, or a vague pointer, "as when one judges 'I will know it when I experience it again'" (2003: 118). We can grant that phenomenal concepts incorporate phenomenal tokens, but could we explain how phenomenal concepts get their referents fixed, just in terms of these associated phenomenal tokens? I think we cannot: these tokens are not sufficient for fixing the reference of the concepts. This is due to the fact that a single phenomenal token instantiates *several* phenomenal properties. For instance, a particular experience of a specific shade of blue is a token of this specific phenomenal quality but also a token of the more general property of being a blue experience. So one phenomenal token cannot determine what phenomenal quality is the referent of each phenomenal

concept. Plausibly, we can have two different phenomenal concepts that incorporate the same phenomenal token: the general concept 'experience of blue' and the more specific concept 'experience of this shade of blue' could be associated with the same phenomenal token, but it is clear that the first concept refers to all experiences of seeing blue and the second refers only to the experiences of seeing that shade of blue. How is this possible?

We can explain this fact if we assume that phenomenal concepts are associated with dispositions to recognise instances of a particular phenomenal kind. As Loar says, "a maximally specific blue phenomenal-token could stand for that maximally specific "shade" of blue only… by way of its association with a disposition to re-identify that "shade" in other phenomenal tokens" (Loar (2003: 119)). That is, our more specific phenomenal concepts refer to more specific phenomenal kinds because they are associated with more specific recognitional dispositions, whereas our more general phenomenal concepts refer to more general phenomenal kinds because of their association with more general recognitional abilities. So what determines that a particular phenomenal concept refers to a particular phenomenal property is that the concept is associated with a certain recognitional disposition to re-identify tokens of that particular phenomenal property.

We can put this idea by saying that phenomenal concepts are *individuated* by our recognitional dispositions. We have as many phenomenal concepts as we have dispositions to re-identify phenomenal types. There are some phenomenal types that are so specific that we are not able to re-identify new instances of them: then, we lack a phenomenal concept for those phenomenal kinds. But in many cases, we do have dispositions to re-identify new tokens of a certain phenomenal kind: these dispositions individuate our phenomenal concepts. [115]

This reference-fixing mechanism is quite *exceptional*. There are other recognitional concepts that are not phenomenal concepts, such as the concept CRAMP, which refers to a kind of muscle contraction by virtue of the property of producing cramp-feelings: the concept refers to the kind of muscle contraction that

[115] Loar (2003) discusses the case of phenomenal properties that are so specific so that we are unable to re-identify them. In these cases, he argues that we do not have phenomenal concepts for these phenomenal properties, since we do not have recognitional dispositions to discriminate them. Loar argues in that paper that this is not a problem for his reply to conceivability arguments, since conceivability arguments can get off the ground only with respect to phenomenal properties for which we have phenomenal concepts. (We need phenomenal concepts in order to articulate conceivability intuitions.)

produces cramp-feelings. (The theoretical concept SUCH AND SUCH KIND OF MUSCLE CONTRACTION refers to the same phenomenon, but it is not associated to that recognitional ability.) But the concept CRAMP differs from, say, the phenomenal concept PAIN in that the first refers to cramps by virtue of a contingent property of them, namely, causing such and such experiences, whereas PAIN refers to pains by virtue of an essential property of pain, namely, being hurtful.

So, what happens when we apply Loar's account of phenomenal concepts to, say, the identity statement (3)? According to Loar, the account would predict that (3) is a posteriori, even if 'pain' and 'C-fibre firing' are associated with the same primary intension. 'Pain' refers to pain by virtue of a certain recognitional ability, whereas 'C-fibre firing' refers to pain because it is the kind that satisfies certain theoretical description. This difference in the corresponding reference-fixing mechanisms motivates the view that (3) is going to be a posteriori, even if the terms are coreferential. The conclusion is, therefore, that sentences such as (3) can be true even if they are posteriori. Or in other words, these sentences can be necessary, even if conceivably false. Therefore, the conceivability arguments above are not correct.

How does this explanation work exactly? How could we explain the fact that (3) is a posteriori by invoking the difference in reference-fixing mechanisms? Loar's explanation here is a bit sketchy, but I think that the main idea is this: the recognitional character of phenomenal concepts entails that phenomenal concepts fix their referents *directly*, that is, without any conceptual mediation, and therefore, phenomenal concepts are not a priori connected with other physical concepts (even if they are coreferential). [116]

How can phenomenal concepts directly refer to their referents? The answer, as we have seen, is that phenomenal concepts pick out their referents by virtue of the associated recognitional abilities. What determines that a particular phenomenal concept refers to a particular phenomenal property is that the concept is associated with certain recognitional disposition to re-identify tokens of that particular phenomenal property. Crucially, these dispositions discriminate phenomenal properties directly, without the need of any further description of such properties. So the phenomenal concepts that these dispositions individuate refer to phenomenal properties directly, without remainder. In Loar's words:

[116] Loar elaborates on this idea in (1990: 87-88) and elsewhere.

> A recognitional concept may involve the ability to class together, to discriminate, things that have a given objective property. Say that if a recognitional concept is related thus to a property, the property *triggers* applications of the concept. Then the property that triggers the concept is the semantic value or reference of the concept; the concept directly refers to the property, unmediated by a higher-order reference-fixer. (Loar (1990: 87-8))

How does this view about the reference of phenomenal concepts account for the fact that the relevant psychophysical identities are a posteriori? Answer: phenomenal concepts being recognitional concepts entails that we can possess phenomenal concepts without having any *conceptual link* to any physical concept. As Loar says, a recognitional concept "directly refers to the property, unmediated by a higher-order reference-fixer" (*Ibid*). This explains how we can entertain both a physical concept such as C-FIBRE FIRING and a phenomenal concept such as PAIN without them being cognitively tied in any interesting way, even if they are coreferential.

Similarly, Hill and McLaughlin (1999) have argued that the deployment of phenomenal concepts involves a different psychological faculty from that involved in the deployment of physical concepts, and because of this, phenomenal and physical concepts are not a priori connected. According to them, these two different psychological faculties have to do with the different ways in which the reference of physical and phenomenal concepts is fixed. As they say, "if a subject fixes the reference of two concepts in different ways, then it will generally be *a priori* possible to analyze this difference purely in terms of a difference between the psychological states that are involved in the two reference-fixings" (1999: 453). We can then appeal to those different psychological states in order to explain why physical and phenomenal concepts are not a priori connected, so that we obtain a *psychological* explanation of the lack of a priori connection between physical and phenomenal concepts. The two different reference-fixing mechanisms that Hill & McLaughlin cite are the following:

> It is plausible, we maintain, that the reference of the concept of pain is fixed by the fact that subjects have a commitment (or a disposition) to apply the concept to internal states that are experienced directly as having a certain qualitative feel. Further, it is plausible that the reference of (say) the concept of C fibre stimulation is fixed by a stipulation involving a description of the form

> "the neural process that has such-and-such a structure and that is responsible for such-and-such experimental effects in the actual world." Under the assumption that the reference of the two concepts in question is fixed in these very different ways, we can account for the fact that it is impossible to see *a priori* that the concepts have the same reference in purely psychological terms. (Hill & McLaughlin (1999: 453))

Here we have a purely psychological explanation of the fact that we cannot see a priori that a certain phenomenal concept and a certain physical concept are co-referential. In this way, we can construct general explanations of why physical and phenomenal concepts are not a priori connected. [117]

Along similar lines, as we have seen, Loar argues that there is an alternative explanation of the aposteriority of (3) and other identity claims involving phenomenal and physical concepts. A crucial idea is that this explanation is *psychological*: "Concepts are related *a posteriori* if they are not *cognitively* tied in a certain way. By that I mean something psychological, and not semantic or modal" (Loar (2003: 116)). He claims it is not necessary to suppose that two concepts are associated with different properties in order to explain why they are conceptually independent.[118] We can offer an alternative explanation by appealing to the kind of *psychological* entity these concepts are. He says:

> The conceptual independence of phenomenal concepts and physical-theoretical concepts has a deep *psychological* explanation that does not depend on distinctness of expressed properties. Suppose (*per impossibile*, if you prefer) that phenomenal concepts refer to physical properties and have the very properties they refer to as their own modes of presentation. Those concepts will themselves be very different *psychological entities* from the theoretical physical-functional concepts that also express those very properties. The latter concepts are realized […] in a verbal-theoretical part of the brain while phenomenal concepts are realized in a nonverbal-experiential part of the brain. That explains their cognitive separation. (Loar (1999: 169))

[117] The claim is that phenomenal concepts are not a priori connected with physical concepts, but this is compatible with there being conceptual ties among phenomenal concepts.

[118] Loar frames the question at issue here in terms of whether any two terms that are not a priori connected must be associated with (or *express*) two different *properties*. However, I have been putting the question in terms of whether any two terms that are not a priori connected must be associated with different *primary intensions*, as we explained in section 3.1. This subtle difference in formulation does not matter too much for our purposes here, since Loar's ideas can also be used to respond to my formulation of the conceivability argument, as we will see.

Loar claims that being *a posteriori* is a consequence of a *psychological* feature, feature which can be explained in terms of the psychological properties of the concepts involved. In the case of the pair of concepts PAIN and C-FIBRE FIRING, they are different kinds of concepts, realized in different parts of the brain, and they play different conceptual roles, so that they are not directly cognitively tied. This "should […] suffice for one concept's […] not being directly inferable from the other without further premises, and hence being related *a posteriori*" (Loar (2003: 116)).

So Loar is offering an alternative explanation of the a posteriori status of psychophysical identities to the one provided by the distinct primary-intension model (DPIM). The main idea is that if there is no direct cognitive connection between two concepts, then they are related a posteriori; and in the case of phenomenal concepts and physical concepts, this absence of direct cognitive connection has a psychological explanation: the concepts are realized in a way so that they are not cognitively tied, independently of whether they express the same properties or not (or in our terms, independently of whether the two terms have the same primary intension or not). Therefore, the fact that two terms are not a priori coreferential does not mean that they necessarily express different primary intensions. So (DPIM) is not correct. And likewise, the fact that a statement is a posteriori does not necessarily entail that its primary intension is not necessary (witness (3), for example). Therefore, (CP) is not correct either.

5.4. Stoljar's Objections to the Recognitional Account

In an important recent paper, Daniel Stoljar (2005b) has argued that the version of the phenomenal concept strategy based upon the recognitional account of phenomenal concepts does not work.[119] However, I think Stoljar's arguments do not work. I will argue that Stoljar has misrepresented the resources of the recognitional account, and that the corresponding version of the phenomenal concept strategy, when it is properly understood, does in fact accomplish the required explanatory task.

Stoljar (2005b) discusses different versions of the phenomenal concept strategy: some versions are theories about the *possession* conditions for phenomenal concepts, that is, these theories specify what is necessary in order to possess a

[119] In the next chapter we will examine some additional arguments by Stoljar which are supposed to refute the phenomenal concept strategy in general.

phenomenal concept; other versions can be seen as theories about the *application* conditions for phenomenal concepts. The upshot of both sort of theories is that, since phenomenal concepts have the features in question, as specified in the possession or application conditions, then the psychophysical conditional P→Q turns out to be a posteriori, independently of whether it is necessary or not.

Stoljar offers criticisms of both types of accounts.[120] He argues that they fail because they do not *entail* and therefore cannot *explain* that the psychophysical conditional is a posteriori. In the following section we will see how Stoljar's argument is supposed to work.

5.4.1. Theories of possession conditions for phenomenal concepts

Stoljar begins by presenting a simplified formulation of the concept-possession variety of the strategy, which on his view has clear problems. He then argues that more sophisticated formulations advocated by real philosophers also suffer the same, or similar, problems.

This simple version of the phenomenal concept strategy is what he calls the Experience Thesis:

> *Experience Thesis*: *S* possesses the (phenomenal) concept *C* of experience *E* only if *S* has actually had experience *E*. (Stoljar (2005b: 471))

The structure of this version of the phenomenal concept strategy is the following: "according to the strategy, if the experience thesis is true, the psychophysical conditional is a posteriori" (2005b: 477). That is, this version of the strategy claims that, if phenomenal concepts have the possession conditions specified in the experience thesis, then it follows that P→Q is a posteriori. According to Stoljar the same structure also applies to other, most sophisticated versions: they will just replace the experience thesis with the thesis they find more appropriate.

We will first examine the problems facing the simple version and then we will turn to a more sophisticated, real example.

Stoljar distinguishes between a sentence's being *a priori* and a sentence's being *a priori synthesizable*. A sentence is *a priori* when "a sufficiently logically

[120] He also discusses and criticises other versions of the strategy, such as concept-acquisition versions, but I will focus on the two mentioned above because they seem to me to be the most promising ones.

acute person who possessed only the concepts required to understand it, is in a position to know that it is true" (2005b: 478). A sentence is *a priori synthesizable* when "a sufficiently logically acute person who possessed only the concepts required to understand *its antecedent*, is in a position to know that it is true" (478).

Stoljar argues that the experience thesis does not entail that the psychophysical conditional is not *a priori*. What it rather entails is that the conditional is not *a priori synthesizable*. The experience thesis says that in order to possess a phenomenal concept, we must have had an experience of the kind that this phenomenal concept refers to. Then, it is possible for a logically acute person to possess the concepts required to understand the antecedent of P→Q, that is, the physical concepts, without possessing all the phenomenal concepts required to understand Q. This person would not be in a position to know that P→Q is true, and therefore, the conditional is not a priori synthesizable.

But, according to Stoljar, the fact that the conditional is not a priori synthesizable does not imply that it is not a priori. For there are conditionals that are a priori but not a priori synthesizable. Stoljar argues for this point by means of the following example:

(5) If *x* is rectangular, then *x* has some property or other. (478)

Stoljar claims that this conditional is clearly a priori, since a (logically acute) subject who understands all the words involved would be in a position to know it is true. But there could be a (logically acute) subject who possessed only the concepts required to understand the antecedent, without thereby being able to understand the consequent. Someone could have the concept RECTANGULAR without having the concept PROPERTY. So this conditional is a priori but not a priori synthesizable.

Therefore, Stoljar maintains that the experience thesis does not have any bearing upon the a posteriori character of P→Q. For all the experience thesis entails, (namely, that the conditional is a not a priori synthesizable), the conditional could still be a priori. Stoljar concludes that this version of the phenomenal concept strategy does not work.

What about other versions, such as Loar's account of phenomenal concepts as recognitional concepts, which we presented earlier? This version, according to Stoljar,

is also a thesis about the possession conditions for phenomenal concepts, which he puts as follows:

> *Recognition Thesis*: *S* possesses the (phenomenal) concept *C* of experience *E* only if *S* has certain dispositions to recognize, discriminate and identify *E* if *S* has or undergoes *E*. (476)

According to the recognition thesis, in order to possess a phenomenal concept *C*, we have to have the disposition to recognize phenomenal states that fall under *C* when we have experiences of that kind.

Stoljar argues that this new version of the concept-possession variety of the phenomenal concept strategy does not fare better than the experience thesis. He thinks that all the concept-possession versions suffer in effect the same problem, which he states as follows:

> According to all accounts of concept possession that might plausibly be appealed to by the phenomenal concept strategy, there is some important link between having a concept and having a particular experience. Since it is plausible to say that there is no such link in the case of physical concepts, there is always the possibility of arguing that the psychophysical conditional is not a priori synthesizable. But the problem is that this makes no advance on the central point, viz., how to move from something's not being a priori synthesizable to something's not being a priori. A different account of concept possession will not help with this problem. (2005b: 481)

Stoljar's claim then is that the problem for the recognition thesis is closely related to the problem for the experience thesis, and it is a general problem for all concept-possession versions: the main idea of all these versions is that 'there is some important link between having a concept and having a particular experience'. However, for Stoljar this only explains why P→Q is not *a priori synthesizable*, not why it is *a posteriori*. These theories about the possession conditions of phenomenal concepts explain how we can possess all physical concepts without thereby possessing all phenomenal concepts, and in this way they explain how we can understand P without understanding Q, hence without being able to know that P→Q is true. But the theory that we need is one that explains why P→Q is a posteriori for subjects that possess all the relevant concepts, including phenomenal concepts. And

according to Stoljar, accounts of the possession conditions of phenomenal concepts cannot help on that point.

Stoljar claims that the phenomenal concept strategy has the obligation to offer a thesis about phenomenal concepts such that it *entails* that P→Q is a posteriori, while at the same time being compatible with the claim that such conditional is necessary. He argues that Loar's thesis about recognitional concepts does not entail that the psychophysical conditional is a posteriori; it only entails that it is not a priori synthesizable.

Stoljar's key assumption is that all concept-possession versions of the phenomenal concept strategy attempt to explain the aposteriority of P→Q by invoking the link between having a phenomenal concept and having an experience. I will argue that this assumption is wrong. Even if we accept that for the recognition thesis (as a concept-possession version of the strategy), there is some important link between possessing a phenomenal concept and having some particular experience, it does not follow that any account employing the recognition thesis has to explain the aposteriority of P→Q in terms of such a link.

As we have seen in the previous section, the crucial element of the recognition thesis is that phenomenal concepts are recognitional concepts. That is, not only is it the case that possessing phenomenal concepts is related somehow to having had certain experiences, but also, and more crucially, possessing phenomenal concepts is related to having certain *recognitional dispositions*. Then, versions of the phenomenal concept strategy based upon the recognition thesis can formulate explanations of the a posteriori status of the psychophysical conditional in terms of such recognitional dispositions, explanations which go beyond the mere contention that there is an important link between having phenomenal concepts and having had the corresponding experiences.

In my view, the thesis that "there is some important link between having a concept and having a particular experience" (Stoljar (2005b: 481)) is *not* the relevant aspect of the recognition thesis for the purposes of the phenomenal concept strategy. As I understand the phenomenal concept strategy, we can use the recognition thesis in order to motivate what I will call *the psychological distinction thesis*. And this second thesis is what explains that P→Q is a posteriori.

In what follows, I will first explain what the psychological distinction thesis is. I will then show how Loar uses it in order to explain the aposteriority of the

psychophysical conditional. Finally, I will show how the recognition thesis can motivate the psychological distinction thesis.

5.4.2. The psychological distinction thesis

> *Psychological distinction thesis:* Phenomenal concepts and physical concepts are psychologically distinct and, in particular, the two types of concepts fail to be cognitively tied in certain ways.

Loar (2003) formulates a version of this thesis and makes clear that he appeals to it in order to explain the a posteriori status of P→Q:

> So the question is this: how might we explain the a posteriori status of a psychophysical property identity? (…) There is a down to earth non-semantic explanation. Concepts are related *a posteriori* if they are not *cognitively* tied in a certain way. By that I mean something psychological […]. The ordinary pain concept we may suppose to have a special connection with neural pain centers; and the theoretical concept of c-fiber stimulation a special connection with verbal theoretical centers. It would be hardly surprising if neural relations do not support a direct cognitive connection. (Loar (2003: 116-117))

The psychological distinction thesis claims that phenomenal concepts and physical concepts are different kinds of psychological entities that are not directly cognitively tied. This implies, crucially, that even when we have had the relevant experiences, and therefore possess *all* the corresponding phenomenal concepts, these are still not cognitively tied to physical concepts. Thus, if the psychological distinction thesis has any consequence with respect to the epistemic status of the conditional P→Q, it cannot be just that the conditional is not *a priori synthesizable*. For this is a matter of what subjects who possess only the concepts needed to understand the antecedent can know a priori. But it is clear that the psychological distinction thesis concerns subjects who possess the concepts needed to understand *both* the antecedent and the consequent of the psychophysical conditional, that is, both physical and phenomenal concepts.

My contention is that the psychological distinction thesis is what explains why psychophysical identity claims such as 'pain is C-fibre stimulation' are a posteriori,

and consequently, why the conditional P→Q is a posteriori too.[121] In order to spell this out, we have to examine what notions of apriority and aposteriority are being invoked here.

According to Loar, "the conceptual property of *being a posteriori* must be in part psychological-cognitive, in a non-semantic sense. It is then hard to see why that psychological relation, or lack of one, should not suffice for one's concept being or not being directly inferable from the other without further premises, and hence being related a posteriori. If some such idea is adequate, it would seem to undercut the idea that we need something contingent in the semantics to explain the *a posteriori* status of phenomenal-physical identities" (Loar (2003: 116)). The main idea here then is that a sentence's being a posteriori can be explained without invoking a contingent proposition expressed by it. The sentence's being a posteriori can have a purely psychological explanation, in terms of the constituent concepts not being cognitively related in a way that makes some inferable from the others.

Once it is clarified what notion of aposteriority is relevant here, it is not very difficult to explain the a posteriori status of psychophysical identity claims such as 'pain is C-fibre stimulation'. Psychophysical identities of this form are a posteriori *because* the corresponding concepts are not psychologically connected in the appropriate way, as the psychological distinction thesis states. Therefore, we can see that the *psychological distinction thesis* can explain the aposteriority of psychophysical identities, and consequently, that of psychophysical conditionals (rather than merely explaining their non a priori synthesizability, as Stoljar suggested).

A further question is: what role does *the recognition thesis* play in explaining the aposteriority of those psychophysical identities claims? My suggestion is that the view that phenomenal concepts are recognitional concepts is what supports the psychological distinction thesis. Thus the view about phenomenal concepts being recognitional plays an indirect role in the explanation of the aposteriority of P→Q, since it serves to support a separate thesis (the psychological distinction thesis) that in turn entails that the conditional is a posteriori.

[121] It is important to notice that Loar's defence of the possibility that the conditional is both necessary and a posteriori proceeds by defending first the possibility that psychophysical identities (involving phenomenal concepts and physical-functional concepts) such as 'pain is C-fibre firing' are both necessary and a posteriori. If psychophysical identities of this sort are necessary and a posteriori, then P→Q will be necessary and a posteriori too.

It should now be clear that the recognition thesis does not have to explain the aposteriority of P→Q by appealing to the fact that there could be subjects who possess all physical concepts but not phenomenal concepts (as Stoljar suggested). What the recognition thesis does is to motivate the claim that physical concepts and phenomenal concepts are different kinds of psychological entities, so that they are not cognitively tied in a way that makes some inferable from the others a priori. Therefore, a logically acute subject that possessed all relevant concepts, that is, *both* the physical and phenomenal concepts, would not be in a position to know that P→Q is true, because physical concepts and phenomenal concepts are not psychologically connected in the appropriate way

According to Loar, as I explained in 5.3, the radically different cognitive roles of the two sorts of concepts are responsible for the fact that the concepts remain unconnected a priori. In the case of phenomenal concepts, the special reference-fixing mechanism that is at work (namely, the recognitional abilities associated with phenomenal concepts) can explain the absence of an a priori connection with physical concepts. As we saw earlier, the crucial idea here is that if phenomenal concepts are recognitional concepts, then phenomenal concepts refer *directly* to their referents, without any sort of mediation or link to other concepts, and therefore, we can possess phenomenal concepts without having any conceptual link to any physical concept. And as we have seen above, this psychological disconnection between physical and phenomenal concepts is enough for explaining why the corresponding identities are a posteriori.

To sum up: the recognition thesis can explain the lack of psychological connection, and this lack of psychological connection can explain the a posteriori status of the relevant propositions. Hence we can conclude that the recognition thesis can explain in this indirect way why the psychophysical conditional remains a posteriori.

It is clear, then, where Stoljar's argument against the concept-possession versions of the phenomenal concept strategy fails: he wrongly assumes that these versions must explain the a posteriori status of P→Q in terms of the possibility of possessing physical concepts without thereby possessing phenomenal concepts. But as I have argued, this is irrelevant for the concept-possession versions based on the recognition thesis. Aposteriority is not explained in terms of subjects missing certain

concepts, but rather in terms of phenomenal and physical concepts missing certain kinds of cognitive ties.

5.5. Conclusion

In this chapter, we have explored the exceptionalist strategy against two-dimensional arguments, and we have focused on the version based on the recognitional account of phenomenal concepts. I have argued that (i) the recognitional account is a plausible account of phenomenal concepts, which provides a plausible explanation of how we can acquire new phenomenal concepts, and how they get their referents fixed (namely, by means of an associated recognitional ability, more or less fine-grained, according to the specificity of the concept). Moreover, I have argued that (ii) if the account of phenomenal concepts as recognitional concepts is correct, then it follows that two terms can be a posteriori connected even if they have the same primary intension, and therefore, (DPIM) would be false under that account of phenomenal concepts. This second claim (ii) shows that the principle (DPIM) cannot be a priori true, and (i) and (ii) jointly entail that (DPIM) is false. In addition, we have argued that, since phenomenal concepts are recognitional, it follows that P&~Q could be both conceivable and false at all possible worlds considered as actual, and therefore, principle (CP) would be false under such an account of phenomenal concepts. This shows that the principle (CP) is not true either. Finally, we have explored Stoljar's important objections against this version of the phenomenal concept strategy, and I have argued that they do not work. Stoljar's main objection had it that the recognitional version of the phenomenal concept strategy could explain only why the psychophysical conditional P→Q is not a priori synthesizable, not why it is not a priori. In response, I have argued that Stoljar's interpretation of the concept-possession variety of the phenomenal concept strategy is wrong: this version of the strategy does not attempt to explain the a posteriori character of P→Q in terms of the fact that we can possess physical concepts without phenomenal concepts, as Stoljar claimed. Rather, that version of the strategy offers an explanation of such aposteriority in terms of what I called the *psychological distinction thesis*, which, if true, would clearly entail that the psychophysical conditional is not a priori true, as required by the phenomenal concept strategy.

6. The Exceptionalist Strategy II: Replies to Objections

In the previous chapter we explored the recognitional account of phenomenal concepts and the prospects for a version of the phenomenal concept strategy based on it. I argued that certain objections against the recognitional account were misguided. In this chapter we will examine some other more general objections that have been raised against the exceptionalist (or phenomenal concept) strategy. I will argue that they do not work, and therefore, I will conclude that the exceptionalist strategy is a successful response against conceivability arguments.

6.1. A Priori vs. A Priori Synthesizable Conditionals

Stoljar (2005b) discusses another problem for the recognition-based phenomenal concept strategy, although, as we will see, this objection also affects other possible versions of the strategy. Let's recall here for convenience the recognition thesis:

> *Recognition Thesis*: *S* possesses the (phenomenal) concept *C* of experience *E* only if *S* has certain dispositions to recognize, discriminate and identify *E* if *S* has or undergoes *E*.

Stoljar introduces the problem by considering the following statement:

> (6) If *x* is a number then *x* is not a red sensation. (2005b: 479)

Recall that, according to the recognition-based version of the phenomenal concept strategy, the recognition thesis can explain the aposteriority of P→Q. However, Stoljar argues that *if* the recognition thesis entailed that P→Q is a posteriori, it would also entail that (6) is a posteriori. In particular, if possessing the concept RED SENSATION requires certain recognitional abilities, possessing NOT A RED SENSATION requires those recognitional abilities too (among other things).[122] Therefore, both RED SENSATION and NOT A RED SENSATION seem to satisfy the recognition thesis. Then, if the recognition thesis entailed that P→Q is a posteriori,

[122] For possessing NOT A RED SENSATION requires possessing RED SENSATION.

it would also entail that (6) is a posteriori as well. But Stoljar claims that (6) is a priori true. So the recognition thesis does not entail that (6) is a posteriori, given that (6) is *not* a posteriori (and, we are assuming, (6) satisfies the recognition thesis in the appropriate way). It follows that the recognition thesis does not entail that P→Q is a posteriori either.

More generally, we can put the problem as follows. It seems that the recognitional version of the phenomenal concept strategy is committed to something like the following claim:

> (PCS_{RT}) For any concepts C_1 and C_2, if C_2, unlike C_1, satisfies the recognition thesis, then 'If *x* is C_1 then *x* is C_2' is not a priori true.

The problem is that there seem to be counterexamples to this principle, such as (6). For NOT A RED SENSATION, unlike NUMBER, satisfies the recognition thesis, but (6) is nonetheless a priori true.

Stoljar also discusses this problem in relation to a version of the strategy based on the *application conditions* for phenomenal concepts. Hill and McLaughlin (1999) and Sturgeon (2000) present versions of the strategy according to which our use of phenomenal concepts and our use of physical concepts are governed by different "epistemic constraints". That is, the experiences that guide us in applying the concepts are different in each case. Phenomenal concepts are *self-presenting concepts*, in the following sense: 'it is a conceptual truth that if I have a red sensation, and if I have the concepts and focus my attention on the matter, I will thereby come to know that I am having a red sensation' (Stoljar (2005b: 483)). We can call this view *the self-presenting thesis*. Physical concepts are not self-presenting in this sense. According to this version of the strategy, this difference between physical concepts and phenomenal concepts entails that the conditional P→Q is a posteriori.

Stoljar says it is plausible to assume that phenomenal concepts are self-presenting concepts. What is not plausible, according to him, is the claim that this feature of phenomenal concepts is responsible for the conditional P→Q being a posteriori. The problem here is, again, that if the self-presenting thesis were to entail that P→Q is a posteriori, then it would also entail that (6) is a posteriori.

Stoljar argues that if RED SENSATION is a self-presenting concept, then NOT A RED SENSATION is self-presenting as well. For in order to apply a self-

presenting concept correctly, we just have to be in the state the concept refers to. We can apply RED SENSATION correctly just by virtue of having a red sensation, and we can apply NOT A RED SENSATION correctly just by virtue of not having a red sensation.

Hence, if phenomenal concepts being self-presenting entailed that P→Q is a posteriori, then NOT A RED SENSATION being self-presenting would entail that (6) is a posteriori. That is, if conditionals with non self-presenting concepts in the antecedent and self-presenting concepts in the consequent, such as P→Q, must be a posteriori, then (6) must be a posteriori too. But (6) is clearly a priori, Stoljar argues, so the self-presenting character of phenomenal concepts cannot be the explanation of the aposteriority of P→Q either.

What should we make of these new arguments from Stoljar based on (6)? In my view, it is true that the provided counterexample makes (PCS_{RT}) false, but fortunately the phenomenal concept strategy does not have to be committed to (PCS_{RT}). More generally, it is not necessary for the success of the strategy that *all* conditionals whose consequents (unlike the antecedents) satisfy the corresponding account of phenomenal concepts be a posteriori conditionals. The strategy, properly understood, is *compatible* with there being some conditionals that satisfy the corresponding possession conditions or application conditions, but are nonetheless a priori, for additional reasons that are not present in the case of the psychophysical conditional. If this correct, then conditionals such as (6) do not really pose a problem for versions of the phenomenal concept strategy based on the recognition thesis or the self-presenting thesis (nor, for that matter, on the experience thesis).

Let's examine now some cases of conditionals that satisfy the recognition thesis in the appropriate way, and nonetheless seem to be a priori true. I will argue that an advocate of the phenomenal concept strategy can accept all the intuitions supporting those counterexamples, without putting the strategy in jeopardy.

For instance, it seems clear that if we possess the concepts NUMBER and SENSATION then we can see that any sentence of the form 'if x is a number then x is a not a sensation of kind K', where K stands for any phenomenal kind, is going to be true. We can know that any sentence of this form is going to be true, even for cases of sensations that we have not had. I can suppose that there is a shade of red, call it 'red*', that I have never seen. If so, I do not possess the concept NOT A RED* SENSATION. But it seems plausible to say that I know a priori that the sentence 'if x

is a number, then x is a not a red* sensation' will express a true proposition, even if I am not able to completely grasp that proposition, due to my failure to grasp one of its constituent concepts.

There are other cases of sentences whose concepts in the consequent, unlike those in the antecedent, satisfy the recognition thesis (and the self-presenting thesis), but where the conditionals seem a priori true. One such example is the following:

(7) If *x* is a square circle, *x* is a red sensation

In this case, the sentence is a priori true because the antecedent of the conditional is a priori false. The main concept in the consequent, RED SENSATION, unlike the main one in the antecedent, satisfies the recognition thesis and the self-presenting thesis, but this conditional is a priori nonetheless.

Do these cases pose a problem for the claim that the recognition thesis (or the self-presenting thesis) explains the aposteriority of the psychophysical conditional? I do not think so. In cases such as (6) and (7), if we grant that these conditionals are a priori true[123], it can be argued that we have *general reasons* for thinking they are a priori true. Any sentence of the form 'if *x* is a number then *x* is not a sensation of kind K' seems to be a priori true. Furthermore, the sentence 'if *x* is a number then *x* is not a sensation' seems to be a priori true too. Any sentence of the form 'if *x* is a square circle, then *p*' is a priori true, for any proposition *p*.

The phenomenal concept strategy does not have to assume that every conditional whose consequent involves concepts with the features specified in the recognition (or self-presenting) thesis will be a posteriori. The strategy is compatible with the claim that there are conditionals of that sort that are a priori, for *additional* reasons. The possession conditions (or application conditions) of phenomenal concepts will entail that certain conditionals are a posteriori *only* when these additional reasons are not present.

The crucial point is that there can be general reasons for thinking that (6) and (7) are not a posteriori, compatible with P→Q being a posteriori. As we have seen, (7) is a priori true because the antecedent is a priori false: any conditional with such an antecedent will be a priori true. And in the case of (6), we could explain why it is a

[123] I am assuming that these sentences are a priori, for the sake of discussion. The denial of this assumption would constitute another way of replying to Stoljar, which I will not pursue here.

priori true in the following way. For instance, we could believe that the following principles hold a priori:

(*Number*): If *x* is a number, then *x* is an abstract entity.
(*Mental*): If *x* is a mental state, then *x* is not an abstract entity.
(*Sensation*): If *x* is a red sensation, then *x* is a mental state.

If these three principles hold a priori, then it follows that (6) is a priori true.[124] I am not defending the view that these principles really *are* a priori. Rather, my contention is that an advocate of the phenomenal concept strategy *could* claim that the three principles are a priori, consistently with her account of phenomenal concepts. In particular, she could hold that those three principles are a priori true, and this would be compatible with the view that her account of phenomenal concepts entails that P→Q is a posteriori.

The moral to be drawn from this discussion is the following: the phenomenal concept strategy cannot be committed to the strong claim that *for every* concept satisfying the recognition thesis (or the self-presenting thesis), it is the case that it is not a priori connected to *any* physical or theoretical concept. However, I think that the phenomenal concept strategy does not have to be committed to that claim. All that the phenomenal concept strategy needs to say is that phenomenal concepts are not a priori connected with the physical-theoretical concepts that appear in the *antecedent* of the psychophysical conditional, that is, P→Q. The strategy does not need to show that *no* phenomenal concept is a priori connected to any physical-theoretical concept. If there are plausible reasons for thinking that some phenomenal concept is a priori connected with some physical-theoretical concept, why should the advocate of phenomenal concept strategy not accept that?

What the strategy needs in order to block conceivability arguments is *just* an account of phenomenal concepts that entails that the *psychophysical conditional* P→Q is a posteriori. This explanation can be as complex as needed. If we can find an account of phenomenal concepts that would make the conditional a posteriori, that is all we need. Imagine that we have an account of phenomenal concepts such that most

[124] This is why: '*x* is a number' a priori entails '*x* is an abstract entity', by (Number) and Modus Ponens. '*x* is an abstract entity' a priori entails '*x* is not a mental state', by (Mental) and Modus Tollens. '*x* is not a mental state' a priori entails '*x* is not a red sensation', by (Sensation) and Modus Tollens. So '*x* is a number' a priori entails '*x* is not a red sensation'. That is, (6) is a priori true.

phenomenal concepts are not a priori connected to physical-theoretical concepts, with the exception of some special cases, such as sentences (6) and (7), which are held to be a priori for the reasons just explained.[125] If such an account of phenomenal concepts *was true*, the psychophysical conditional would be *a posteriori*. And such an account is perfectly compatible with the recognitional account, or the self-presenting thesis. Therefore, Stoljar's objection does not really pose a serious problem to the claim that the recognition thesis or the self-presenting thesis, properly understood, can explain why P→Q is a posteriori.

6.2. The Conceivability Argument against Behaviourism

I would like to consider one final argument from Stoljar (2005b) against the phenomenal concept strategy. The main idea of his argument is the following: if the phenomenal concept strategy was a successful strategy against conceivability arguments against physicalism, it would also be a successful strategy against conceivability arguments against *behaviourism*. But, he claims, these arguments against behaviourism are good arguments, so the phenomenal concept strategy has to be wrong. Let's look at how Stoljar's argument is supposed to work in more detail.

Stoljar characterises *supervenient behaviourism* as "the thesis that the phenomenal truths supervene with metaphysical necessity on the behavioural truths" (484). Supervenient behaviourism, thus characterised, entails that the *psychobehavioural conditional* (that is, B→Q, where B is a description of all facts about actual and potential behaviour, and Q is any phenomenal truth) is *necessarily* true. Stoljar then rehearses a conceivability argument against supervenient behaviourism, similar in structure to the conceivability argument against physicalism discussed above.[126] The argument goes like this:

---

[125] Some other special cases could be for instance: 'If *x* is an ethical principle, then *x* is not a toothache', or 'If *x* is law of nature, then *x* is not a sensation of heat'. If we accept that (6) is a priori true, we could also accept that these are a priori true. The advocate of the phenomenal concept strategy could easily explain why these are a priori, along the same lines as the explanation concerning (6): we could use the principle (Mental) and other appropriate principles. For instance, we could say that '*x* is an ethical principle' a priori entails '*x* is abstract', which a priori entails '*x* is not a toothache', and so on.

[126] Stoljar claims that this conceivability argument against behaviourism has been widely used in philosophy of mind. About this, see the discussion below.

> I can conceive of someone behaviourally identical to me and yet who lacks experiences altogether (…). But this is just to say that I can conceive of a situation in which the psychobehavioural conditional is false. And of course, if this is so, then [B→Q] is at best contingent and supervenient behaviourism is false. (485)

Stoljar "take[s] as a datum that this argument is a good one" (485). He then goes on to argue that if the phenomenal concept strategy was a successful response to the conceivability argument against *physicalism*, by parity of reason it would also be a successful response to the conceivability argument against *behaviourism*. That is, if our favourite version of the phenomenal concept strategy can explain why P→Q is a posteriori, the same strategy would explain why B→Q is a posteriori.

As we have seen, the phenomenal concept strategy argues that P→Q is *a posteriori*, due to certain features of phenomenal concepts, and therefore independently of whether such a conditional is necessary or not. Accordingly, the conditional can be both a posteriori and necessary. If this response was successful, the advocate of supervenient behaviourism could respond as follows: "phenomenal concepts are self-presenting [or recognitional, or whatever feature of phenomenal concepts we take to explain the a posteriori status of the psychophysical conditional], while behavioural concepts are not. Given these differences between phenomenal concepts and behavioural concepts, a phenomenal state and its nomologically correlated behavioural state would seem contingently related, even if they were necessarily one" (485). In this way, identities between phenomenal and behavioural kinds could be both necessary and a posteriori, which would explain why B→Q is necessary and a posteriori.

For Stoljar, it is obvious that this cannot be a good response on behalf of supervenient behaviourism, since he is assuming that no response against the conceivability argument against supervenient behaviourism above can be good. So he infers that the phenomenal concept strategy's response against the conceivability argument against physicalism is not successful either.

In response, I think that there are several problems with this argument from Stoljar. First, I think that the argument against behaviourism that Stoljar presents in his paper is different from the classical arguments against behaviourism suggested by Putnam and Block, among others, and which are widely endorsed in contemporary philosophy of mind. Secondly, I think that Stoljar is not entitled to assume that the

conceivability argument against behaviourism that he presents is a *good* argument against supervenient behaviourism and that therefore there can be no successful response against it. It is question-begging to take that for granted, particularly in the context of this discussion about conceivability arguments and responses to them. Thirdly, I think that the conceivability argument that Stoljar endorses is just wrong, for the same considerations that make the conceivability argument against physicalism wrong. And fourthly, I think that there are alternative arguments against supervenient behaviourism that do not have the structure of conceivability arguments, which can be (and indeed have been) used to defeat supervenient behaviourism, and this makes even clearer that we have no good reasons to accept Stoljar's conceivability argument against supervenient behaviourism.

Before examining each of these points in turn, let us reformulate Stoljar's proposed argument against behaviourism a bit more precisely. As we have seen, the argument presented above infers the possibility of B&~Q from its mere conceivability. We can reformulate the argument as follows:

The Conceivability Argument against Behaviourism

B1. B&~Q is conceivable.

B2. If B&~Q is conceivable, then it is possible.

B3. So B&~Q is possible.

B4. If B&~Q is possible, supervenient behaviourism is false.

———————

B5. Supervenient behaviourism is false.

The first problem with Stoljar's argument is the following: there certainly is a type of argument against behaviourism in the literature, which has the structure of a conceivability argument and has been widely accepted. However, this type of argument is *not* the same as the argument that Stoljar rehearses, because the conclusion is rather different. Classical arguments against behaviourism are not directed against *supervenient* behaviourism, but rather against the stronger thesis of *logical behaviourism*. This is so, for instance, in Hilary Putnam's "Brains and Behaviour", and in Ned Block's "Psychologism and Behaviorism".[127] That is, the

[127] See Block (1981) and Putnam (1968).

arguments put forward in these influential papers are mainly addressed to the view that "there exist entailments between mind-statements and behaviour-statements (…) that follow from the meanings of mind words. I shall call these *analytic entailments*" (Putnam (1968: 46)). As Block (1981) puts it, "behaviourist analyses were generally presented as *conceptual* analyses" (30), and therefore the relevant counterexamples involve *conceptual* claims. That is, if logical behaviourism takes the conditional B→Q to be a conceptual truth, then we can offer counterexamples against it, in terms of conceivable situations that satisfy B but not Q. Block offers such an example in the form of a complex machine with complex behaviour but no intelligence.[128] It seems clear that if B&~P is conceivable, then B→Q cannot be a conceptual or analytic truth. This is what the classical arguments against logical behaviourism assert. And the phenomenal concept strategy is perfectly compatible with this point. But the argument (B1-B5) above is much stronger than this, and therefore the classical anti-behaviourist arguments do not offer any support for it.

So in conclusion, I think that Stoljar is simply wrong to suppose that his conceivability argument against supervenient behaviourism is widely endorsed in the philosophy of mind literature. Rather, a related argument against logical behaviourism—which in no way can be used to argue against the phenomenal concept strategy—is the one that is widely endorsed.

Secondly, I think that it is question-begging in this context to assume that the argument against supervenient behaviourism suggested by Stoljar is a good argument, independently of whether it has been widely endorsed or not. I think it is not supported in the background literature, but moreover it begs a crucial question in this debate: whether the conceivability of a sentence such as B&~Q entails its possibility. Or in other words: whether a conditional such as B→Q can be both necessary and a posteriori. The advocates of conceivability arguments claim that conceivability does entail possibility (in the relevant sense), and therefore a conditional of that sort cannot be both necessary and a posteriori. The advocates of the phenomenal concept strategy try to reject this, by appealing to the views on phenomenal concepts seen above. In this debate, one cannot just take for granted an argument whose premises involve the claim that the conceivability of a similar statement (B&~Q) entails its possibility.

[128] Block's original argument was concerned with the supervenience of intentional truths and intelligence, rather than with our present issue, namely, phenomenal truths. But the counterexamples could be easily modified to be applied to the question of consciousness.

In order to see this point more clearly, we can distinguish two formulations of the conceivability argument against behaviourism suggested by Stoljar. One is just the version above, that is, (B1-B5). A second formulation is the following:

The Conceivability Argument against Behaviourism II

B1. B&~Q is conceivable.

(CP) If S is conceivable, then S is possible.

B2. If B&~Q is conceivable, then B&~Q is possible.

B3. So B&~Q is possible.

B4. If B&~Q is possible, supervenient behaviourism is false.

———————

B5. Supervenient behaviourism is false.

In this second formulation, the principle that warrants the inference from B1 to B2 is made explicit, namely, the familiar conceivability-possibility thesis that we have been discussing so far. My point against Stoljar can be put in the form of a dilemma: if we understand the argument against supervenient behaviourism in the first way, then it is no problem for the phenomenal concept strategy to accept that it is a sound argument. And if, on the other hand, we understand the argument against supervenient behaviourism in the second way, then the phenomenal concept strategy cannot hold that it is a sound argument, but it would clearly be question-begging for Stoljar to assume it is. Let me elaborate both horns in turn.

First, it is easy to see that the advocates of the phenomenal concept strategy can claim that the first argument (B1-B5) is a *sound* argument. For they can endorse all the premises: they can (and typically do) accept that the conclusion is true, that is, supervenient behaviourism is false. Crucially, they can accept that B3 is true, namely, that B&~Q is possible. If so, then it follows that the conditional B2 is true too, because its consequent is just B3, and it is clear that if B3 is true, the conditional B2 will be true as well; and they can easily accept that B1 is true too. Thus, the advocates of the phenomenal concept strategy can accept that the first argument is sound. The crucial point is that this is perfectly compatible with the phenomenal concept strategy: what the strategy entails is that the conditional B→Q has to be a posteriori, independently of whether it is necessary or not. But the strategy is not committed to the further claim that the conditional is necessary. It is perfectly compatible with the

strategy to claim that the conditional B→Q is both a posteriori and not necessary (since B&~Q is possible). What the strategy does claim is that the fact that the conditional B→Q is a posteriori does not entail that it is not necessary. But this does not mean that it has to be necessary.

However, the fact that the first argument (B1-B5) is a *sound* argument does not mean that it is a *successful* argument against behaviourism. To illustrate this, we can compare it with other sound arguments against supervenient behaviourism. Let's consider this trivial argument:

(i) The sky is blue.
(ii) If the sky is blue, B&~Q is possible.
(iii) B&~Q is possible.
(iv) If B&~Q is possible, supervenient behaviourism is false.
(v) So supervenient behaviourism is false.

If we take (ii) to be a material conditional, then this argument is sound: it is valid and all premises are true. But it is obvious that, even if sound, this is not a *good argument* against supervenient behaviourism. Likewise, the fact that B3 above is true does not automatically make that argument a *good* one, even if we accept that that would make (B2) true. What is needed is independent support for B2.

The second formulation of the conceivability argument against supervenient behaviourism above does attempt to provide independent support for B2, namely, premise (CP). This second formulation is in effect incompatible with the phenomenal concept strategy. But the problem now is that it is very clear that Stoljar is not entitled to assume that this formulation is a good argument because this would imply that principle (CP) is correct, and this is exactly the issue that is at stake here, so Stoljar cannot just assume it. Assuming (CP) as part of an argument against the phenomenal concept strategy is clearly question-begging.

My third point against Stoljar's argument is simply this: if the considerations in the previous chapters are correct, then we have good reasons to believe that (CP) is wrong. Therefore, it is perfectly OK for the advocate of the phenomenal concept strategy to claim that the conceivability argument against supervenient behaviourism II is not correct.

Finally, it is important to bear in mind that there are other ways of arguing for B3 and B5, which do not require (CP). There are other arguments for the claim that supervenient behaviourism is false. It is not necessary at this point to argue that these other arguments are more successful. My contention is just that there are some *alternative* arguments. For instance, one such alternative argument against supervenient behaviourism is the so-called *causal argument* against behaviourism,[129] and it goes like this: supervenient behaviourism claims that there are no phenomenal states over and above behavioural dispositions. But then, how could phenomenal states be causally responsible for those behavioural dispositions? If phenomenal states are just dispositions to behave in certain ways, it would be incorrect to claim that those phenomenal sates are *causally* responsible for the manifestation of such dispositions. For instance, if my itch is identified with a disposition to scratch, then we cannot say that my itch causes my disposition to scratch, because something cannot cause itself. Also, the itch could not be said to cause the scratching itself, because this is just a manifestation of the disposition to scratch, and dispositions do not cause their manifestations.[130] This is a plausible argument against supervenient behaviourism, but it does not depend on the conceivability of B&~P. Rather, this argument relies upon the empirical claim that phenomenal states cause behavioural dispositions. Therefore, the plausible claim that (B3) and (B5) are correct does not justify in any way the conceivability argument against behaviourism II above.

My conclusion is that the phenomenal concept strategy does not face any problem concerning the defeat of behaviourism. The phenomenal concept strategy is compatible with widely accepted claims that are part of the background in contemporary philosophy of mind, namely that *logical* behaviourism has been falsified by conceivability intuitions, and in addition that supervenient behaviourism has also been falsified (where this point does not have to be committed to the claim that only conceptual considerations could have falsified supervenient behaviourism: empirical considerations could very well be responsible for that).

Of course, the phenomenal concept strategy would be in trouble if the argument II above was a good argument. But Stoljar has not offered any good reason to believe that it is a good argument. And if the considerations in previous chapters

[129] See, for instance, Braddon-Mitchell and Jackson (1996: 34-35)
[130] This point is made in Braddon-Mitchell and Jackson (1996: 35).

are correct, and the phenomenal concept strategy is successful, that shows that such an argument is indeed wrong.

6.3. Chalmers' Dilemma for the Phenomenal Concept Strategy

In this section I will present and criticise a final argument against the phenomenal concept strategy, which has been presented by Chalmers (2007). According to Chalmers, any successful version of the phenomenal concept strategy should have the following structure:

> Proponents put forward a thesis C attributing certain psychological features—call these the key features—to human beings. They argue (1) that C is true: humans actually have the key features; (2) that C explains our epistemic situation with regard to consciousness: C explains why we are confronted with the relevant distinctive epistemic gaps; and (3) that C itself can be explained in physical terms: one can (at least in principle) give a materialistically acceptable explanation of how it is that humans have the key features. (2007: 172)

Chalmers argues that while the three elements (1) to (3) are essential for the strategy to work, no account of phenomenal concepts can satisfy both (2) and (3). That is, he argues that no account of phenomenal concepts (call it C) that an advocate of the phenomenal concept strategy might propose can simultaneously satisfy the three requirements that are essential for the strategy to succeed. In particular, no account of phenomenal concepts can both be explicable in physical terms, and strong enough to explain our epistemic situation with respect to consciousness. Therefore, Chalmers concludes that any version of the phenomenal concept strategy is irredeemably condemned to fail in resisting the inference from the epistemic gap to an ontological gap.

We can see Chalmers' argument for this conclusion as presenting a dilemma for the phenomenal concept strategy:

I. If P&~C is conceivable, then C is not physically explicable.
II. If P&~C is not conceivable, then C cannot explain our epistemic situation.

———————

III. Either C is not physically explicable, or C cannot explain our epistemic situation.[131]

The dilemma starts in the following way: for any account of phenomenal concepts C, either it is entailed a priori by a complete physical description of the world (P), or it is not. That is, either the conditional P→C is a priori true or not. If the conditional is a priori true, then P&~C is not conceivable. If the conditional is not a priori true, P&~C will be conceivable.

What does it mean to say that P&~C is conceivable? It means that we are able to imagine a physical duplicate of the world, that is, a world that satisfies P, which does not satisfy the strategy's account of phenomenal concepts, that is, C. If we can imagine such a world, then P does not entail C a priori, and therefore P&~C would be conceivable. In particular, we can focus on zombie worlds: physically identical duplicates of this world where no-one is conscious. According to the proponent of the phenomenal concept strategy, zombie worlds are conceivable (even if they are not possible). Zombie worlds would clearly satisfy P. So if we could imagine zombie worlds that do not satisfy C, then P&~C would be conceivable. If, on the other hand, we could not imagine a physically identical world where C does not hold, not even a zombie-world, then P&~C would not be conceivable. These two options represent the two horns of Chalmers' dilemma, that is, the two premises of his argument. I will examine these two premises in turn.

The *first premise* explores what happens if P&~C is conceivable. According to Chalmers, the problem for the phenomenal concept strategy in this case is that P does not entail C a priori, and this has the consequence that we cannot *explain* that account of phenomenal concepts in physical terms, given Chalmers' sense of explanation. According to Chalmers, a reductive explanation of X in physical terms requires that P entails X a priori. Therefore, P has to entail C a priori if we want a reductive explanation of our account of phenomenal concepts in physical terms.

Chalmers has argued that in order to explain some macro-phenomenon in microphysical terms, we have to show how microphysical facts give rise to that macro-phenomenon, and this requires that microphysical facts transparently show how the macro-phenomenon obtains.[132] If microphysical facts did not entail a priori

[131] Chalmers (2007: 174).
[132] We explained this model of reductive explanation, i.e. the *functional model*, in section 3.4.

the instantiation of the macro-phenomenon, then we would require some additional explanatory materials in order to explain such an instantiation, and this would show that the original explanation was incomplete.

Someone could argue that the phenomenal concept strategy should not accept this much. For if we assume that reductive explanation requires a priori entailment, then it is clear that, if there is an epistemic gap between P and Q (that is, no a priori entailment from P to Q), then there is no reductive explanation of consciousness in physical terms. Can an advocate of the phenomenal concept strategy accept this claim? In my view, it is perfectly coherent (although by no means compulsory) for the phenomenal concept strategy to accept Chalmers' account of reductive explanation, given that what is at issue here is whether the phenomenal concept strategy can defend the thesis of *physicalism* from conceivability arguments, not so much the defence of the possibility of a reductive explanation of consciousness.[133] So for the sake of discussion I will assume that reductive explanation of X in physical terms requires that P entails a priori X.[134]

Someone might also wonder why it is necessary for the phenomenal concept strategy to offer a reductive explanation of their proposed account of phenomenal concepts in physical terms. According to Chalmers, in order for the phenomenal concept strategy to work, it has to show that physicalism is compatible with the epistemic gaps, and therefore, its proponents have to show how physical facts alone give rise to the key features of phenomenal concepts which in turn give rise to the relevant epistemic gaps. If they do not provide a physically acceptable explanation of

[133] Some versions of the phenomenal concept strategy are indeed concerned with the defence of the possibility of a reductive explanation of consciousness in physical terms (e.g. see Carruthers (2004) and Tye (1999)), but the version of the strategy that Chalmers (2007) mainly focuses on is just concerned with the issue of physicalism. I will follow him in this respect, focusing on the question of whether the phenomenal concept strategy can block the conceivability argument as an argument against physicalism.

[134] Such assumption, even if compatible with the phenomenal concept strategy, is indeed controversial, for independent reasons. I am making such an assumption here for the sake of discussion, but for the record, I do not really accept it. One way of arguing against Chalmers' account of reductive explanation is the following: we could use the claim that we argued for in Ch. 4, namely, that microphysical truths do not entail ordinary macroscopic truths a priori, for arguing that reductive explanations of ordinary macroscopic truths in physical terms do not really require a priori entailments. For it is widely accepted that we can have reductive explanations of ordinary non-phenomenal macroscopic truths in physical terms, but then if these macroscopic truths are not a priori entailed by microphysical truths, this shows that reductive explanation does not require a priori entailment. For some compelling arguments against Chalmers' account of reductive explanation along these lines, see Block & Stalnaker (1999).

the key features of phenomenal concepts, then they will not have motivated the view that their explanation of the epistemic gap is indeed compatible with physicalism.

The *second premise* of the argument examines the possibility that P&~C is *not* conceivable, that is, that a complete physical description of the world (P) a priori entails the corresponding account of phenomenal concepts C. In this case, even zombie worlds satisfy C: every conceivable physical duplicate of the actual world satisfies C. This horn of the dilemma represents the possibility that we are really capable of offering a physically acceptable explanation of the key features of phenomenal concepts (according to Chalmers' standards). So we would not have the problem of the previous horn.

Unfortunately, Chalmers argues, this second horn of the dilemma faces another problem: if our account of phenomenal concepts is tame enough to be physically explicable (in Chalmers' sense), then it could not explain our epistemic gap. Let me elaborate why this is so, according to Chalmers, in some detail.

If P entails C a priori, then even zombies satisfy C. But Chalmers argues (and this is the crucial point) that zombies do *not share our epistemic situation* with regards to consciousness. If so, then C does not entail a priori that someone is in our epistemic situation with regard to consciousness, because we can conceive of some beings (zombies) that satisfy C but do not share our epistemic situation. And therefore, C cannot explain our epistemic situation (again, assuming Chalmers' account of explanation as a priori entailment).

Why does Chalmers think that zombies would not share our epistemic situation? Well, zombies seem to have a very different epistemic position with regards to consciousness from ours. We are conscious, but they are not. We have true beliefs about our own phenomenal states, but they do not. We have knowledge of phenomenal facts, but they do not. Our epistemic position is very different from the epistemic position of zombies. Chalmers characterizes someone's epistemic situation as including "the truth values of their beliefs and the epistemic status of their beliefs (as justified or unjustified, and as cognitively significant or insignificant)" (2007: 176). He assumes we can draw a correspondence between our beliefs and zombies' beliefs. Then, he adds: "A zombie will share the epistemic situation of a conscious being if the zombie and the conscious being have corresponding beliefs, all of which have corresponding truth-value and epistemic status" (177). So in order to evaluate whether zombies share our epistemic situation we have to compare the truth-values

and epistemic status of our beliefs and their corresponding beliefs. If there is any difference, our epistemic positions differ.

Chalmers considers some of zombies' beliefs such as 'I am phenomenally conscious'. He suggests that this belief does not seem to be true of zombies. This could seem a very trivial point, since zombies are not conscious by definition. But actually this issue turns out to be more complicated, since the truth-values of the zombies' corresponding beliefs depend on what their contents are, and it is not clear what their contents could be. Since zombies lack consciousness, how could their corresponding beliefs be about consciousness?

Chalmers realizes that judgements about the truth-value of the zombies' corresponding "phenomenal" beliefs are controversial, and therefore he does not want to rely on them. So he focuses on the issue of the justification of the zombies' corresponding "phenomenal" beliefs. In order to do this, he uses Jackson's example of Mary (the colour-scientist who had spent all her life in a black-and-white room, until one day she was released and was able to see a red object for the first time).[135] He also introduces a new character, Zombie-Mary, that is, a physical duplicate of Mary who does not have any conscious mental states. Chalmers asks us to consider the following two utterances, by Mary and Zombie-Mary respectively:

Mary (after being released): 'I am having a red sensation'
Zombie-Mary (after being released): 'I am having a red sensation'

Chalmers argues that the beliefs expressed by these utterances are not equally justified. Mary's judgement is justified by her own phenomenal states, and arguably she gains new substantive phenomenal knowledge by having that sensation for the first time. Zombie-Mary's judgement does not have the same kind of justification, since she is not enjoying any kind of phenomenal state. Maybe she is gaining some new sort of knowledge (indexical, for instance), but she is not gaining new *phenomenal* knowledge. So, according to Chalmers, zombies' beliefs have a different justification status, and therefore, they do not share our epistemic situation, even if they satisfy C. Hence, C cannot explain our epistemic situation.

[135] See section 3.3 above.

Chalmers concludes that for any account of phenomenal concepts, either it cannot be physically explicable (first horn of the dilemma) or it cannot explain our epistemic situation with regards to consciousness (second horn). This, then, is Chalmers' new response to the phenomenal concept strategy. In the remainder of this chapter, I will argue that it fails.

6.4. The Second Horn of the Dilemma: Explaining the Epistemic Gap

In my response to Chalmers, I will focus on the second horn of the dilemma. I think that, even if we agree with Chalmers' argument that zombies' "phenomenal" beliefs have a different justificatory status from ours, Chalmers' second horn of the dilemma does not pose a serious challenge for the phenomenal concept strategy. The problem is that Chalmers has misidentified the explanatory goal of the phenomenal concept strategy. In my view, what an account of phenomenal concepts has to explain is simply the epistemic gap between physical and phenomenal truths, not our *entire* epistemic situation with regards to consciousness. And, as I will argue, Chalmers has presented no obstacle to the idea that an account of phenomenal concepts might explain the epistemic gap in that sense.[136]

I think it is clear, from our presentation of the conceivability argument and the phenomenal concept strategy, that what C should explain is the *epistemic gap* between the physical and the phenomenal, because it is precisely this epistemic gap which is supposed to imply an ontological gap. The aim of the phenomenal concept strategy is to show that there is an alternative explanation of the epistemic gap in terms of certain features of phenomenal concepts, an explanation that does not entail that there is an ontological gap. The strategy is not supposed to explain *our full epistemic access* to consciousness.

[136] One might also challenge Chalmers' assumption that zombies "phenomenal" beliefs have a different justificatory status from ours. For instance, Carruthers and Veillet (forthcoming) defend the view that zombies do share our epistemic situation. They argue that zombies' corresponding "phenomenal" beliefs can be both true and justified, they just have a different content: they are about zombies' quasi-phenomenal states. However, I am not going to base my defence of the phenomenal concept strategy on the view that zombies' "phenomenal" beliefs are true and justified. Instead, I will focus on the view that what the strategy has to explain is the epistemic gap between physical and phenomenal beliefs, which can be characterised independently of those phenomenal beliefs' truth and justification. Nevertheless, if it did turn out to be the case that zombies' corresponding beliefs are true and justified, that would be compatible with my own arguments against Chalmers' dilemma.

The question at issue, then, is the following: how should we characterize the epistemic gap between physical and phenomenal truths that the phenomenal concept strategy aims to explain? In 6.4.1, I will introduce the characterization of the epistemic gap that I favour, and I will argue that zombies do not pose a problem for the task of explaining the epistemic gap in that sense. Secondly, in 6.4.2 I will explain Chalmers' alternative understanding of the epistemic gap to be explained by the strategy. I will then argue that an advocate of the phenomenal concept strategy has no good reason to accept Chalmers' characterization.

6.4.1. The epistemic gap, properly understood

My suggestion is that all that C has to explain is the fact that P (the complete physical description of the world) does not entail Q (any phenomenal truth) a priori. I think that the phenomenal concept strategy does not have to explain our whole epistemic situation but just why there is an *inferential disconnection* between physical truths and phenomenal truths, that is, why the latter are not a priori inferable from the former.

So the relevant question, then, if we want to know whether an account of phenomenal concepts C entails a priori our epistemic gap, is the following: can we imagine beings that satisfy C but fail to satisfy the inferential disconnection between P and Q? But, what is it for a being to *fail* to satisfy that inferential disconnection? Well, since in order to satisfy the inferential disconnection, a subject has to be such that she is *not* able to infer Q a priori from P, it seems that she will fail to instantiate such a disconnection when, in effect, she is *able* to infer Q a priori from P.

Chalmers argues that zombies do not share our epistemic gap. If we understand the epistemic gap as an inferential disconnection between P and Q, then the question at stake would be this: can we imagine zombies that satisfy C *and* are able to infer Q a priori from P?

We can easily see that there cannot be such zombies. The reason is that zombies cannot even entertain Q, since they do not possess the concepts required to understand Q, namely, phenomenal concepts. They do not have phenomenal states, which are essential in order to possess (most of) our phenomenal concepts, so they cannot have phenomenal beliefs (with the same content as ours). It follows that zombies cannot entertain Q, and therefore, zombies are *not able* to infer Q from P a priori. For in order to infer Q a priori from P, first you need to entertain Q. So zombies *do satisfy* the inferential disconnection between P and Q. Hence, they do

satisfy the epistemic gap in the sense I am advocating. Therefore, there is no reason to think that C does not entail a priori the epistemic gap.

Chalmers seems to agree with the claim that zombies cannot entertain Q: "It is plausible that a nonconscious being such as a zombie cannot have beliefs with exactly the same content as our beliefs about consciousness" (2007: 177). They can have corresponding beliefs but with different content: in the case of our phenomenal beliefs, zombies can have corresponding *quasi-phenomenal* beliefs (about their corresponding quasi-phenomenal states).

Therefore, Chalmers could reply that what is at issue here is rather whether zombies would be able to infer a priori their *corresponding* "phenomenal" belief Q* from their corresponding belief P. So let's examine this question: can zombies a priori infer Q* from P? I think that they cannot. The reason is the following. Recall that we are considering the second horn of Chalmers' dilemma, according to which P&~C is not conceivable, and therefore even zombies satisfy C. Then, zombies' corresponding phenomenal concepts satisfy the account of phenomenal concepts C, so those phenomenal concepts are not a priori connected to physical concepts. So P will not entail Q* a priori.

We can see this point a bit more clearly by focusing on a particular account of phenomenal concepts. For instance, we can consider Hill & McLaughlin's account (explained in Ch. 5) according to which phenomenal concepts and physical concepts play very different psychological roles, and this is what explains the lack of a priori connection. We could characterise these psychological roles in purely *functional* terms, and therefore a zombie (that is, a functional duplicate of us) would also have concepts that played those different roles. So we could talk about zombies' corresponding quasi-phenomenal concepts (those concepts that are functionally equivalent to our phenomenal concepts) and zombies' corresponding physical concepts. Since these zombie-concepts also play different roles, they will not be a priori connected, and therefore, sentences involving quasi-phenomenal concepts cannot be a priori inferred from sentences involving only physical concepts.[137]

Therefore, we can conclude that, if we understand the epistemic gap as an inferential disconnection between physical and phenomenal beliefs, then there is no

[137] As we have seen earlier, different versions of the phenomenal concept strategy explain this lack of a priori connection in different ways, but they all agree that phenomenal concepts are such that the relevant sentences involving them are not a priori inferable from sentences involving physical concepts only.

evidence that C might hold without the epistemic gap holding. In particular, Chalmers' alleged counterexample, namely, zombies, is not a real case of beings that satisfy C but not the epistemic gap, since there is no *relevant* epistemic gap that they fail to satisfy, even if they do not instantiate all the aspects of our epistemic situation. Hence, zombies do not represent a problem for the claim that C can explain the epistemic gap.

6.4.2. Chalmers' reply: the knowledge-involving epistemic gap

Chalmers seems to accept that there is some kind of inferential disconnection that is also present in the case of zombies, but he claims that to explain the inferential disconnection in that sense is not enough for explaining our epistemic gap. He agrees that zombies are not able to entertain Q, which entails that they are not able to infer Q from P a priori, and he also agrees that they are not able to a priori infer Q* from P either: "it is plausible that a zombie's physical and quasi-phenomenal beliefs are no more inferentially connected than a conscious being's beliefs" (2007: 183). Still, he thinks that zombies fail to instantiate the *relevant* aspects of *our* epistemic gap. Let's see why:

> Whereas the inferential disconnection strategy might physically explain an inferential disconnection between physical and phenomenal *beliefs*, the anti-physicalist's crucial epistemic gap involves a disconnection between physical and phenomenal *knowledge*. (2007: 183-4)

> In the anti-physicalist's arguments, the relevant epistemic gap (from which an ontological gap is inferred) is characterised in a way that truth and knowledge are essential. […] It is crucial to the conceivability argument that one can conceive beings that lack phenomenal states that one actually has. And it is crucial to the explanatory gap that one has cognitively significant knowledge of the states that we cannot explain. (2007: 183)

What Chalmers is claiming at this point is that the inferential disconnection that the phenomenal concept strategy should explain is a disconnection between physical truths and phenomenal beliefs that are *true and justified*. In other words, C has to explain how we can conceive of beings that are physically identical to us but lack the phenomenal states of which *we have* substantive knowledge. For Chalmers, I take it, a subject has substantive knowledge of phenomenal states if and only if she has

phenomenal states. Therefore, Chalmers is claiming here that an account of phenomenal concepts has to explain the epistemic gap in the following sense: a subject will instantiate the epistemic gap when (a) she has phenomenal states (so that she can have justified true beliefs about them) and (b) she cannot infer a priori her phenomenal beliefs from a physical description of the world.

However, I want to argue that C merely has to entail a weaker version of the epistemic gap, which includes (b), but not necessarily (a). That is, in order for a subject to instantiate the epistemic gap in this weaker sense, she must instantiate the appropriate inferential disconnection between P and Q, but it is not necessary for her to have genuine phenomenal states, nor believe that she has them, nor know that she has them.

I think that the demand for an explanation of the epistemic gap in the stronger sense is not motivated. In order to show this, I will first argue (in 6.4.2.1) that there is a straightforward problem with Chalmers' claim that the phenomenal concept strategy should explain the epistemic gap in the strong sense (i.e. both (a) and (b)): this is clearly too strong, because if we accept Chalmers' characterization of the epistemic gap, then it will follow that in order for C to explain the epistemic gap, the epistemic gap must be closed. Secondly (in 6.4.2.2) I will examine the argument that Chalmers presents for the claim that the phenomenal concept strategy should explain the epistemic gap in the strong sense, and I will argue that it does not provide sufficient motivation.

6.4.2.1. Closing the epistemic gap?

Recall that Q can be any phenomenal truth, so let Q be a proposition of the form 'I am phenomenally conscious'. And recall that Chalmers' characterization of the epistemic gap requires (a), that is, the subject must have phenomenal states, so her corresponding Q has to be true. Then, we can characterise this epistemic gap as follows:

Chalmers' Epistemic Gap (E): Q & It is not a priori that (P→Q)

I take it that this characterization is a priori true. For the epistemic gap has been *defined* as crucially involving phenomenal knowledge, and, plausibly, the having of phenomenal knowledge a priori entails the having of phenomenal states. That is, E

entails Q a priori. And remember that, on this horn of the dilemma, C is a priori entailed by P. So, if we assume that C a priori entails E, this will follow:

P *a priori entails* C
C *a priori entails* E
E *a priori entails* Q
_____
Therefore, P *a priori entails* Q

This would mean that there is no epistemic gap between P and Q. That is, if we characterize the epistemic gap as E, then, on this horn of the dilemma, in order for an account of phenomenal concepts to explain the epistemic gap, there cannot be an epistemic gap!

This is problematic, for the following reason. Chalmers himself claims, elsewhere in the paper, that we should characterize both C (the account of the key psychological features of phenomenal concepts) and the *epistemic situation* to be explained by C in "*topic-neutral* terms: terms that do not explicitly attribute phenomenal states or concepts that refer to them. (…) This allows the possibility that even if consciousness cannot be physically explained, we might be able to physically explain the key psychological features and our epistemic situation" (2007: 175). This sounds very plausible.

However, I think that it is clear that Chalmers' characterization of the epistemic gap does not allow the possibility that even if consciousness cannot be physically explained, the epistemic gap could still be explained. As we have seen above, if we characterize the epistemic gap as E, then for C to entail E a priori, it has to be the case that P entails Q a priori. But if P does not entail Q a priori (that is, if consciousness cannot be physically explained) then C cannot entail E a priori either, and therefore C cannot explain E (in Chalmers' sense of explanation). So Chalmers' characterization of the epistemic gap has the consequence that, if consciousness is not physically explainable, the epistemic gap will not be physically explainable either. In other words: Chalmers' characterization of E is not really *topic-neutral*, because it

involves phenomenal states, as we can see above. Therefore, it is a bad characterization.[138]

My conclusion is that we should reject Chalmers' characterization of the epistemic gap as E, and we should endorse a characterization that really is topic-neutral. My characterization of the epistemic gap as an inferential disconnection between physical and phenomenal beliefs does satisfy that constraint, and therefore there are good reasons to prefer it. And as I have shown earlier, versions of the phenomenal concept strategy that take the second horn of the dilemma, that is, that offer accounts of phenomenal concepts that are physically explicable, do not have problems in explaining the epistemic gap in the preferred sense.

6.4.2.2. The conceivability argument and the epistemic gap

Given these difficulties, why does Chalmers hold that we have to explain the epistemic gap in the strongest sense, that is, one that involves the presence of phenomenal states themselves? Chalmers says that the epistemic gap in that sense (the truth-and-knowledge-involving epistemic gap) is the one which is supposed to entail an ontological gap, according to the conceivability argument. Therefore, he argues, it is the epistemic gap in that sense that the phenomenal concept strategy has to explain. He adds: "If one characterised these gaps in a way that were neutral on the truth of phenomenal beliefs, the arguments would not get off the ground" (2007: 183). Why should we believe that the truth of our phenomenal beliefs is crucial for getting the

[138] In discussion, Chalmers has suggested that, on his view, it is not true *by definition* that satisfying the epistemic gap requires having phenomenal states: as a matter of fact, the epistemic gap requires having phenomenal states, but this is not part of the definition of epistemic gap. That is, he denies that E entails Q *a priori*. However, he says in his paper that "In the anti-physicalist's arguments, the relevant epistemic gap (from which an ontological gap is inferred) is characterised in a way that truth and knowledge are essential. […] It is crucial to the conceivability argument that one can conceive beings that lack phenomenal states that one actually has." (2007: 183) Here, the epistemic gap does seem to require having *phenomenal states* (by definition), and therefore, this is not a topic-neutral characterization of the epistemic gap. More importantly, though, if we drop the requirement that E entails Q a priori, then it is not clear what Chalmers' response to my argument in 6.3.2. would be. If the epistemic gap just involves the inferential disconnection between P and Q, and not the having of phenomenal states *per se*, then zombies can satisfy the epistemic gap too. Chalmers' argument requires a more substantial characterization of the epistemic gap, such that zombies cannot instantiate the epistemic gap in that sense, but it is topic-neutral. It is not clear what this notion could be. Chalmers' suggestion seems to be to characterise the epistemic gap as the inferential disconnection between physical and phenomenal (or pseudo-phenomenal) beliefs, where the latter are *true and justified*, but where this notion of justification does not involve having phenomenal states by definition, although non-conscious beings cannot, as a matter of fact, have that kind of justification. In response, I think that, first, it is not clear whether such a notion is coherent, and secondly, even if it was, it is not clear what the motivations for preferring it over my own characterization of the epistemic gap are. For more on this second point, see 6.4.2.2.

conceivability argument off the ground? I think that this has to do with the fact that the conceivability argument is a valid argument against physicalism only if at least some phenomenal beliefs are true. Let's review the conceivability argument (in a simplified form, since we can ignore some technical details here) in order to see why:

CA1: P&~Q is conceivable.
CA2: If P&~Q is conceivable, then P&~Q is metaphysically possible.
CA3: If P&~Q is metaphysically possible, physicalism is false.
_____
CA4: Physicalism is false.

As Chalmers argues, premise CA3 is plausible only if there is some phenomenal belief Q that is true of the actual world. That is, physicalism is committed to the claim that any physical duplicate of the world is such that all *truths* about the world are true there as well. If there were no phenomenal truths, then the possibility of P&~Q would not be a problem for physicalism at all.

So we can agree that in order to get the conceivability argument *against physicalism* off the ground, Q has to be true. However, this does not force us to maintain that the *epistemic gap* (premise CA1) requires the truth of some phenomenal belief Q. Recall that the phenomenal concept strategy's aim is to block the inference from the conceivability of P&~Q to its possibility, that is, they want to dispute CA2. And as I have explained, the premise that requires the truth of some phenomenal belief Q is CA3. That is, the epistemic gap can be formulated independently of whether Q is true or not. Of course, the *argument* itself can get off the ground only if Q is true. But this is irrelevant for the phenomenal concept strategy, whose main aim is to establish that there is an alternative explanation of the epistemic gap that does not have the consequences that CA2 asserts. This task is completely independent of whether CA3 is true or not. Therefore, there is no good reason here to believe that we should characterise the epistemic gap in a sense that involves the truth of Q.

6.5. Conclusion

In the first two sections of this chapter, we explored some important objections against the phenomenal concept strategy from Stoljar, and I argued that they do not work.

The first objection by Stoljar had it that the phenomenal concept strategy, if successful, would have the consequence that some plausibly a priori propositions are not really a priori true. In response, I argued that the phenomenal concept strategy does not have to be so committed.

Stoljar's second objection had it that if the phenomenal concept strategy was successful in defending physicalism from conceivability arguments, it would also be successful in defending behaviourism from analogous conceivability arguments, but, Stoljar assumes, the latter arguments are correct, so the phenomenal concept strategy cannot be successful against the former sort either. In response, I distinguished between being committed to the *falsehood* of behaviourism and being committed to the *conceivability arguments* against behaviourism, and I explained that the phenomenal concept strategy can be committed to the former claim without being committed to the latter. I argued that traditional arguments against behaviourism in the literature are not of the sort Stoljar advocates, and in addition, I argued he is not entitled to assume that his version of the argument against behaviourism is a good argument.

Therefore, we can conclude that Stoljar's objections do not carry any force against the phenomenal concept strategy.

Finally, we examined Chalmers' dilemma for the phenomenal concept strategy. He argues that no matter what horn is taken, there will be serious problems. I replied that, on the contrary, an advocate of the phenomenal concept strategy could take the second horn of the dilemma, and still have a successful response to conceivability arguments. Of course, if she accepts the existence of an explanatory gap between the physical and the phenomenal, then she cannot provide a physical explanation (in Chalmers' sense) of the epistemic situation that *conscious beings* are in. But I argued that this is not relevant at all for the success of the strategy. For it is enough, I believe, that the strategy presents an account of phenomenal concepts which explains why physical truths do not entail phenomenal truths a priori; the account does not have to explain how we can *know* phenomenal truths. Therefore, I think that there is at least one way out of Chalmers' dilemma.

In conclusion, we can indeed assert that the phenomenal concept strategy is very much alive and well.

Conclusion

In this thesis we have discussed several arguments against the claim of physicalism, and we have argued that they do not ultimately work. Here I will briefly summarize these arguments and my responses to them, and then I will draw some final conclusions.

First of all, we got clearer about what physicalism says. We saw that the intuitive claim of physicalism is committed to the claim that any minimal physical duplicate of the actual world is a duplicate concerning all properties. We also saw that the intuitive claim of physicalism is not committed to the claim that every property instantiated in the actual world is identical to some physical property.

We then introduced the anti-physicalist arguments that we have been critically assessing in this thesis: the conceivability arguments against physicalism. These arguments attempt to establish that physicalism is false, merely by *conceptual* grounds, and in particular, by focusing on what we can conceive. We focused on two of the arguments that seemed initially more compelling: the two-dimensional argument against physicalism and the two-dimensional argument against type-identities. We noticed that this second argument is not strictly speaking an argument against physicalism, but we also noticed that both arguments depend on an influential semantic framework, namely, two-dimensional semantics. In particular, we identified a core claim of two-dimensional semantics that both arguments rely on, namely, (2D): 'S is a priori if and only if S is 1-necessary (that is, true in all possible worlds considered as actual)'. Hence, this principle became our main target. By criticising this principle, we showed that conceivability arguments cannot falsify either physicalism or psychophysical identities merely on conceptual grounds.

We examined two different strategies against (2D). According to the first one (the *non-exceptionalist* strategy), this principle fails across the board. According to the second one (the *exceptionalist* strategy) this principle fails when it comes to sentences involving phenomenal concepts. We elaborated, developed and defended both sorts of strategies.

Our discussion of the non-exceptionalist strategy focused on certain conditionals which have been submitted as direct counterexamples to (2D), such as the conditional 'If physical facts are such and such, then facts about water are such

and such'. We explored Chalmers and Jackson's detailed discussion of such counterexamples. They argue that since these conditionals are ultimately a priori true, then they cannot constitute a counterexample to (2D). We examined the view about concepts and concept-possession that they use in order to support their argument (which we called 'reductive ascriptivism'), and we concluded that this view is problematic.

However, we saw that, even if there were some way of defending reductive ascriptivism, the principle (2D) would still be in jeopardy. For we saw that there is another strategy against (2D), namely, the exceptionalist one, which does not rely on reductive ascriptivism being wrong. In particular, we saw that, even if (2D) turned out to be true when applied to sentences involving ordinary non-phenomenal macroscopic concepts, this does not entail that (2D) will be true when applied to sentences involving *phenomenal* concepts. This strategy argues that sentences such as 'If physical facts are so and so, then phenomenal facts are so and so' would be a posteriori even if they were 1-necessary. If this is so, then it is clear that (2D) is false, since the fact that a sentence is not a priori does not entail that it is not 1-necessary.

As we saw, the main idea of this strategy is that phenomenal concepts have some special features that make them not inferable a priori from physical concepts, even if they were to refer to physical properties. Then it is clear that a conditional of the sort 'If physical facts are so and so, then phenomenal facts are so and so' would not be a priori, given the special features of phenomenal concepts, but that conditional could very well be 1-necessary. Again, this shows that (2D) is false. In order to show this, we do not have to show that the conditional 'If physical facts are so and so, then phenomenal facts are so and so' is indeed a counterexample to (2D), in the sense that it is in fact a posteriori and 1-necessary. The argument here is rather the following: according to our account of phenomenal concepts, it follows that, if (per impossible, if you prefer) such conditional turned out to be 1-necessary, it would still not be a priori true. And this is enough to show that (2D) is false.

Therefore, we can conclude that two-dimensional arguments are wrong on at least two counts. On the one hand, they rely on a certain view about (ordinary non-phenomenal macroscopic) concepts, namely, reductive ascriptivism, which is very problematic. (In particular, as we explained in Ch. 4, reductive ascriptivism has it that mere possession of such concepts would entitle someone to infer truths involving such concepts from a microscopic description of the world, which, we argued, does not

seem to be the case.) On the other hand, we saw that, even if those views about concepts turned out to be correct concerning non-phenomenal macroscopic concepts, we have good reasons to believe that they do not apply to phenomenal concepts. In particular, we examined a very plausible account of phenomenal concepts, namely, the *recognitional* account, and we argued that this account entails that (2D) is not correct.

These two strategies have been presented here as two different ways of criticizing conceivability arguments that are incompatible (although complementary). As we have seen, the non-exceptionalist strategy *denies* reductive ascriptivism, whereas the exceptionalist strategy typically *endorses* reductive ascriptivism (with respect to macroscopic non-phenomenal concepts). However, there is nothing in the main ideas behind the exceptionalist strategy that forces us to endorse the view of reductive ascriptivism. That is, one advantage of the exceptionalist strategy is that it can work *even if* reductive ascriptivism is correct. But we can see that it can also work even if this view is ultimately wrong, as the non-exceptionalist strategy has it. Therefore, the core ideas of both strategies are ultimately compatible. Let me elaborate this a little.

We have seen that the main idea of the non-exceptionalist strategy is that we are not really able to infer macroscopic truths about the world from microphysical truths a priori. This is perfectly compatible with the exceptionalist strategy's claim that our phenomenal concepts have some special features that make them not inferable from microphysical concepts, such as being recognitional concepts. As I explained in Ch. 5, the main advocate of the recognitional view, namely, Brian Loar, does believe that we can infer (non-phenomenal) macroscopic truths from microphysical truths a priori, but he argues that phenomenal concepts are exceptional in that truths involving them cannot be so inferred. However, it is perfectly possible to hold these two views: (i) that the recognitional character of phenomenal concepts entails that truths involving them are not inferable a priori from microphysical truths, and (ii) that the possession-conditions of our ordinary non-phenomenal macroscopic concepts are such that they cannot be inferred from microphysical truths a priori, even if they are not recognitional concepts.

Hence, we can see that the fact that someone is an advocate of the non-exceptionalist strategy does not prevent her from endorsing the recognitional account

of phenomenal concepts (or any other similar account), and also the claim that this account can be used to provide additional arguments against (2D).

We saw in Ch. 5 that the exceptionalist strategy has also been labelled 'the phenomenal concept strategy'. This label emphasizes the fact that the main idea of these accounts is that the epistemic gap between physical and phenomenal truths can be explained in terms of some special features of phenomenal concepts (and therefore we do not need to posit an ontological gap in order to explain such an epistemic gap). Then, we could use the label 'phenomenal concept strategy' more generally, to refer to all those strategies that attempt to offer alternative explanations of the epistemic gap in terms of the special features of phenomenal concepts, *regardless* of their views on reductive ascriptivism. Therefore, the phenomenal concept strategy, in this sense, is neutral concerning whether (2D) is correct when applied to non-phenomenal macroscopic concepts or not. That is to say, the phenomenal concept strategy in this sense is compatible both with the non-exceptionalist strategy being true and with it being false.

Therefore, using the labels in this way, we can now put the main conclusion of this thesis as follows: The non-exceptionalist strategy shows that there is something wrong with reductive ascriptivism, and therefore, (2D) fails. However, if someone wanted to bite the bullet and endorse reductive ascriptivism in spite of all its problems, there is another way of showing that (2D) fails anyway, namely, by means of the exceptionalist strategy (which, we can stipulate, assumes that the non-exceptionalist strategy does not work). In any case, I have argued that we have good reasons to endorse *both* the non-exceptionalist strategy and the phenomenal concept strategy (which, we can stipulate, is neutral concerning whether the non-exceptionalist strategy ultimately works or not).

Thus, we have very good reasons to think that conceivability arguments against physicalism do not work. Another question is whether physicalism itself is correct or not. In the first chapter we examined an argument to the effect that physicalism is correct, namely, the causal argument, although we realized that there is much more to be said about whether such argument, or other arguments for physicalism, ultimately work or not. But this discussion will have to wait for another time.

Bibliography

Alter, T. and Walter, S. (2007) (eds.) *Phenomenal Concepts and Phenomenal Knowledge: New Essays on Consciousness and Physicalism*, Oxford: Oxford University Press.

Block, N. (1981) "Psychologism and Behaviorism", *Philosophical Review* 90, pp. 5-43.

Block, N., Flanagan, O. and Güzeldere, G. (1997) (eds.) *The Nature of Consciousness*, MA: MIT Press

Block, N. and Stalnaker, R. (1999) "Conceptual Analysis, Dualism, and the Explanatory Gap", *Philosophical Review* 108, pp. 1-46.

Braddon-Mitchell, D. and Jackson, F. (1996) *Philosophy of Mind and Cognition*, Oxford: Blackwell.

Braun, D. (2006) "Names and Natural Kind Terms" in E. Lepore and B. Smith (eds.) *The Oxford Handbook of Philosophy of Language*, Oxford: Oxford University Press, pp. 490-515.

Byrne, A. (1999) "Cosmic Hermeneutics", *Nous* 33, pp. 347-383.

Carruthers, P. (2004) "Reductive Explanation and the 'Explanatory Gap'", *Canadian Journal of Philosophy* 34, pp. 153-173.

Carruthers, P. and Veillet, B. (forthcoming) "The Phenomenal Concept Strategy", *Journal of Consciousness Studies*. Now available at: http://www.philosophy.umd.edu/Faculty/pcarruthers/Phenomenal%20Concept%20Strategy.htm

Chalmers, D. (1996) *The Conscious Mind*, New York: Oxford University Press.

Chalmers, D. (2002a) "Does Conceivability entail Possibility" in T. S. Gendler and J. Hawthorne (eds.) *Conceivability and Possibility*, Oxford: Oxford University Press, pp. 145-200.

Chalmers, D. (2002b) "On Sense and Intension", *Nous* 36, pp. 135-182.

Chalmers, D. (2002c) (ed.) *Philosophy of Mind: Classical and Contemporary Readings*, New York: Oxford University Press.

Chalmers, D. (2003) "Consciousness and its Place in Nature" in T. Warfield and S. Stich (eds.), pp. 102-142.

Chalmers, D. (2004) “Epistemic Two-Dimensional Semantics”, *Philosophical Studies* 118, pp. 153-226.

Chalmers, D. (2007) “Phenomenal Concepts and the Explanatory Gap” in T. Alter and S. Walter (eds.), pp. 167-194.

Chalmers, D. (forthcoming) “The Two-Dimensional Argument against Materialism” in B. McLaughlin and A. Beckermann (eds.) *The Oxford Handbook of Philosophy of Mind*, Oxford: Oxford University Press. Now available at: http://consc.net/papers/2dargument.html

Chalmers, D. and Jackson, J. (2001) “Conceptual Analysis and Reductive Explanation”, *Philosophical Review* 110, pp. 315-360.

Davidson, D. (1970) “Mental Events” in L. Foster and J. W. Swanson (eds.) *Experience and Theory* , Amherst: University of Massachusetts Press, pp. 79-101.

Davies, M. and Humberstone, I. L. (1980) “Two Notions of Necessity”, *Philosophical Studies* 38, pp. 1-30.

Dennet, D. (1988) “Quining Qualia” in A. Marcel and E. Bisiach (eds.) *Consciousness in Contemporary Science*, Oxford: Oxford University Press. Reprinted in D. Chalmers (2002c), pp. 226-246.

Dowell, J. (2006) “Formulating the Thesis of Physicalism: An Introduction”, *Philosophical Studies* 131, pp. 1-23.

Fodor, J. (1974) “Special Sciences (or: The Disunity of Science as a Working Hypothesis)”, *Synthese* 28, pp. 97-115.

García-Carpintero, M. and Macià, J. (2006) *Two-Dimensional Semantics*, Oxford: Oxford University Press.

Hempel, C. (1969) “Reduction: Ontological and Linguistic Facets” in S. Morgenbesser, P. Suppes and M. White (eds.) *Philosophy, Science, and Method: Essays in Honor of Ernest Nagel*, New York: St. Martin's Press, 179-199.

Hill, C. and McLaughlin, B. (1999) “There Are Fewer Things in Reality that Are Dreamt of in Chalmers’ Philosophy”, *Philosophy and Phenomenological Research* 59, pp. 445-454.

Jackson, F. (1982) “Epiphenomenal Qualia”, *Philosophical Quarterly* 32, pp. 127-136.

Jackson, F. (1986) “What Mary Didn’t Know”, *Journal of Philosophy* 83, pp. 291-295.

Jackson, F. (1994) “Finding the Mind in the Natural World” in R. Casati, B. Smith, and G. White (eds.) *Philosophy and the Cognitive Sciences*, Vienna: Holder-Pichler-Tempsky. Reprinted in Chalmers (2002c), pp. 162-169.

Jackson, F. (1998a) *From Metaphysics to Ethics: A Defence of Conceptual Analysis*, New York: Oxford University Press.

Jackson, F. (1998b) “Reference and Description Revisited”, *Nous* 32, pp. 201-218

Jackson, F. (2003) “Mind and Illusion” in A. O’Hear (ed.) *Minds and Persons* (*Royal Institute of Philosophy Supplements* 53), Cambridge: Cambridge University Press, pp. 251-271.

Kim, J. (1993) *Supervenience and Mind*, Cambridge: Cambridge University Press.

Kim, J. (1998) *Mind in a Natural World*, Cambridge, MA: MIT Press.

Kim, J. (2005) *Physicalism, or Something Near Enough*, Princeton: Princeton University Press.

Kirk, R. (1974) “Zombies vs. Materialists”, *Proceedings of the Aristotelian Society*, Supp. Vol. 48, pp. 135-52.

Kirk, R. (2005) *Zombies and Consciousness*, Oxford: Oxford University Press.

Kripke, S. (1980) *Naming and Necessity*, Harvard University Press. Excerpted in Chalmers (2002c), pp. 329-334.

Levin, J. (1991) “Analytic Functionalism and the Reduction of Phenomenal States”, *Philosophical Studies* 61, pp. 211-238.

Levine, J. (1983) “Materialism and Qualia: The Explanatory Gap”, *Pacific Philosophical Quarterly* 64, pp. 354-361. Reprinted in Chalmers (2002c), pp. 354-359.

Levine, J. (1993) “On Leaving Out What It’s Like” in M. Davies and G. Humphreys (eds.) *Consciousness: Psychological and Philosophical Essays*, Oxford: Blackwell, pp. 121-136.

Levine, J. (2001) *Purple Haze: The Puzzle of Conscious Experience*, Cambridge, MA: MIT Press.

Lewis, D. (1994) “David Lewis: Reduction of Mind” in S. Guttenplan (ed.) *A Companion to the Philosophy of Mind*, Oxford: Blackwell, pp. 412-431.

Loar, B. (1990) “Phenomenal States”, *Philosophical Perspectives* 4, pp. 81-108.

Loar, B. (1997) “Phenomenal States (Revised Edition)” in N. Block, O. Flanagan and G. Güzeldere (eds.), pp. 597-616.

Loar, B. (1999) "David Chalmers' *The Conscious Mind*", *Philosophy and Phenomenological Research* 59, pp. 465-472.

Loar, B. (2003) "Qualia, Properties, Modality", *Philosophical Issues* 13, pp. 113-129.

Ludlow, P., Nagasawa, Y. and Stoljar, D. (2004) (eds.) *There's Something about Mary: Essays on Phenomenal Consciousness and Frank Jackson's Knowledge Argument*, Cambridge, MA: MIT Press.

Macdonald, C, (2004) "Mary Meets Molyneux: The Explanatory Gap and the Individuation of Phenomenal Concepts", *Nous* 38, pp. 503–524

Marcus, E. (2004) "Why Zombies are Inconceivable", *Australasian Journal of Philosophy* 82, pp. 477-490.

McLaughlin, B. (1995) "Varieties of Supervenience" in E. Savellos and U. Yalçin (eds.) *Supervenience: New Essays*, New York: Cambridge University Press, pp. 16-59.

Melnyk, A. (2003) "Physicalism" in T. Warfield and S. Stich (eds.), pp. 65-84.

Montero, B. (2003) "The Epistemic/Ontic Divide", *Philosophy and Phenomenological Research* 66, pp. 404-418.

Nagel, T. (1974) "What is it Like to Be a Bat?", *Philosophical Review* 83, pp. 435-450.

Papineau, D. (2002) *Thinking about Consciousness*, Oxford: Oxford University Press.

Papineau, D. (2007) "Phenomenal and Perceptual Concepts" in T. Alter and S. Walter (eds.), pp. 111-144.

Perry, J. (2001) *Knowledge, Possibility and Consciousness*, Cambridge, MA: MIT Press.

Place, U. T. (1956) "Is Consciousness a Brain Process?", *British Journal of Psychology* 47, pp. 44-50.

Putnam, H. (1967) "Psychological Predicates" in W. H. Capitan and D. D. Merrill (eds.) *Art, Mind and Religion*, Pittsburgh: University of Pittsburgh Press, pp. 37-48.

Putnam, H. (1968) "Brains and Behaviour" in R. Butler (ed.) *Analytical Philosophy: Second Series*, Oxford: Blackwell, pp. 1-19. Reprinted in Chalmers (2002c), pp. 45-54.

Putnam, H. (1975) "The Meaning of 'Meaning'" in K. Gunderson (ed.) *Language, Mind and Knowledge*, Minneapolis: University of Minnesota Press, pp. 131-93.

Smart, J. J. C. (1959) "Sensations and Brain Processes", *Philosophical Review* 68, pp. 141-156. Reprinted in Chalmers (2002c), pp. 60-68.

Stalnaker, R. (1978) "Assertion" in P. Cole (ed.) *Syntax and Semantics: Pragmatics*, Vol. 9, New York: Academic Press, pp. 315-332.

Stalnaker, R. (1996) "Varieties of Supervenience", *Nous* 30, pp. 221-241.

Stanley, J. (1997) "Names and Rigid Designation" in B. Hale and C. Wright (eds.) *A Companion to the Philosophy of Language*, Oxford: Blackwell Press, pp. 555-585.

Stoljar, D. (2001a) "Two Conceptions of the Physical", *Philosophy and Phenomenological Research* 62, pp. 253-281.

Stoljar, D. (2001b) "The Conceivability Argument and Two Conceptions of the Physical", *Nous* 35, pp. 393-413.

Stoljar, D. (2005a) "Physicalism", *The Stanford Encyclopedia of Philosophy (Winter 2005 Edition)*, E. N. Zalta (ed.), URL: http://plato.stanford.edu/archives/win2005/entries/physicalism/

Stoljar, D. (2005b) "Physicalism and Phenomenal Concepts", *Mind and Language* 20, pp. 469-494.

Stoljar, D. (2006) *Ignorance and Imagination: The Epistemic Origin of the Problem of Consciousness*, Oxford: Oxford University Press.

Sturgeon, S. (2000) *Matters of Mind*, London: Routledge.

Tye, M. (1999) "Phenomenal Consciousness: the Explanatory Gap as a Cognitive Illusion", *Mind* 108, pp. 705-725.

Warfield, T. and Stich, S. (2003) (eds.) *Blackwell Guide to Philosophy of Mind*, Oxford: Blackwell

White, S. (1986) "Curse of the Qualia", *Synthese* 68, pp. 333-368. Reprinted in N. Block, O. Flanagan and G. Güzeldere (eds.), pp. 695-717.

White, S. (forthcoming) "Why the Property Dualism Argument Won't Go Away" in G. Bealer and R. Koons (eds.) *The Waning of Materialism: New Essays*, Oxford: Oxford University Press. Now available at: http://www.nyu.edu/gsas/dept/philo/courses/consciousness/papers/WHYPDAW.html

Williamson, T. (2003) "Blind Reasoning II: Understanding and Inference", *Proceedings of the Aristotelian Society*, Supp. Vol. 77, pp. 249–293.

Wilson, J. (2006) “On Characterizing the Physical”, *Philosophical Studies* 131, pp. 61-99.

Worley, S. (2003) “Conceivability, Possibility and Physicalism”, *Analysis* 63, pp. 15-23.

Worley, S. (2006) “Physicalism and the Via Negativa”, *Philosophical Studies* 131, pp. 101-126.