Modal Rationalism and the Metaphysics of Mind

Harry Cleeveley

PhD - Philosophy

King’s College London

2019

Declaration

I certify that this thesis is my own original work, and that the work of others has been appropriately acknowledged. This thesis incorporates some material previously submitted for another degree (MPhilStud, Birkbeck, 2015), but the majority has been produced during the period of registration for the degree for which it is submitted.

Acknowledgements

I am extremely grateful to everyone who has helped me with the development of this thesis. Most of all, I would like to thank my Supervisor, David Papineau, for his wise guidance and advice – and for taking the opposite view to me on most issues. I would also like to express my gratitude to everyone at the Department of Philosophy at King's College London for their support, and to thank numerous philosophers at King's and elsewhere for their help and advice over the years: Jessica Leech, Christopher Hughes, Matthew Parrot, Corine Besson, Charlie Pelling, Keith Hossack, Stacie Friend, Ralph Walker, and Lizzie Fricker.

Abstract

Physicalism and externalism are widely held theories about mind. Physicalism is the thesis that the physical properties of the world are sufficient for consciousness; externalism is the claim that certain external objects are necessary for some mental contents. Although these two theories are independent, they do have some important features in common. I will argue that both are false for the same underlying reason, which is that they entail false claims about the space of possible worlds. Physicalism and externalism can both be formulated in such a way as to rule out certain kinds of possible world. But, in both cases, we can use conceivability arguments to show that the worlds in question are in fact possible, and therefore that these doctrines are false. The most famous such conceivability argument is Chalmers's (2010) two-dimensional argument against physicalism; along the same lines, I will set out a two-dimensional argument against externalism.

These conceivability arguments can only succeed if ideal conceivability entails metaphysical possibility. But does it? Since the work of Kripke (1971), the philosophical orthodoxy is that it does not. This is because it is widely held that there are *a posteriori* necessities – that is, there are propositions that are true in every metaphysically possible world, but whose truth cannot be known on the basis of rational reflection alone. The negation of such a truth is therefore conceivable, but not possible. But this view is not universal. Chalmers (2010) argues that the Kripke cases trade on an ambiguity in the referring expressions they contain. Thus they have one sense in which they are necessary, and another sense in which they are *a posteriori,* but no sense in which they are both necessary and *a posteriori.* If this analysis is right, then the Kripke cases do not in fact disprove the entailment from conceivability to metaphysical possibility.

So the fundamental question is whether there are any *strong* necessities – which are necessary and *a posteriori* on the *same interpretation*. Many examples have been proposed; much debate revolves around whether or not these supposed examples are compelling. In this thesis I will argue that there are in fact no strong *a posteriori* necessities at all, because they are impossible – the very idea, in fact, is incoherent. Therefore I conclude, against the conventional wisdom, that ideal conceivability does entail metaphysical possibility. This, in turn, leads to an internalist and anti-materialist metaphysics of mind.

Contents

Introduction

Mind does not sit easily within a naturalistic view of the world. Two mental phenomena especially defy easy explanation. The first is consciousness: there are certain things, such as human beings, that there is *something it is like to be*. We have subjectivity, an inner life. The second is representational content: some mental states involve representing the world as being a certain way – they *mean* something. But if the world is ultimately physical, then how can it contain objects that there is something it is like to be? And how can one thing mean another?

Nonetheless, despite these apparent mysteries, there is a broadly naturalising tendency in the philosophy of mind. There is a widespread view that mental phenomena are ultimately natural phenomena like any other and should be understood as such. In this spirit, two widely-held philosophical views are *physicalism* about consciousness, and *externalism* about mental content. Physicalism is the claim that the material facts of the world are sufficient for consciousness; externalism is the idea that certain external objects are necessary for some mental contents.

In this thesis, I will argue that physicalism about consciousness and externalism about mental content are both false. I will argue that consciousness is not reducible to the physical realm, but rather that it is a fundamental part of reality in its own right. And will I argue that our mental representations do not essentially depend on the existence of external objects, but are constituted by internal facts – indeed, by internal facts about consciousness. Now, this does not necessarily mean that mental phenomena are not natural phenomena. In fact, I think they are. But it does mean that we need to broaden our understanding of the natural world – moving beyond a narrow materialist conception – in order to make room in it for mind.

Moreover, I will not just argue that physicalism and externalism are both false. I will argue that physicalism and externalism, although they are separate claims, are both false for the same underlying reason: they both entail false claims about the space of possible worlds. This type of argument is now familiar when deployed against physicalism, as seen in Chalmers's two-dimensional or zombie argument (1996, 2010). I will consider Chalmers's argument in some detail – I argue that it is ultimately successful – and I will argue that many of the same ideas and principles apply equally to externalism about mental content.

How will these arguments work? The first step is to show that both physicalism and externalism entail that certain kinds of world are impossible. Physicalism, according to the standard formulation (Jackson 1998a), entails that there is no metaphysically possible world that is a minimal physical duplicate of the actual world, but which differs with respect to consciousness. And, as I will show, externalism entails an analogous claim about the space of worlds. Since externalism claims that certain external objects, E, are metaphysically necessary for the occurrence of certain mental contents, M, it follows that if externalism is true, then there is no possible world in which these mental contents occur but the relevant external objects do not.

The next step in the argument is that the worlds described, although supposedly impossible, are perfectly conceivable. We can conceive of a world that is a minimal physical duplicate of the actual world, but which differs with respect to consciousness. And we can conceive of a world where mental contents, M, occur but external objects, E, do not. Then comes the crucial, most controversial step: that *conceivability entails metaphysical possibility.* If all of the above can be proven, then the falsehood of physicalism and externalism follows in short order: each doctrine entails that a certain kind of world is impossible; but the worlds in question are perfectly conceivable; conceivability entails possibility; therefore the worlds in question are possible; therefore the doctrines are false.

Plainly, conceivability arguments such as these can only work if it is legitimate to draw conclusions about metaphysical possibility from the conceivability of a scenario. But does conceivability entail possibility? That is, if we can form a coherent conception of a scenario, does it follow that there is a corresponding metaphysically possible world?

The prevailing view in modern philosophy, following the work of Kripke (1971), is that conceivability does not entail possibility. This is because it is widely held that there are *a posteriori* necessities – that is, that there are some propositions that are true in all metaphysically possible worlds, but which can only be known *a posteriori*. The classic example is that water is H_2O: this cannot be known just by *a priori* reflection – it is conceivably false – and yet it is a necessary truth, or so it is argued, because there is no possible world in which water is not H_2O. If there are indeed *a posteriori* necessities, then their negations will be conceivable true, yet impossible. So the existence of any genuine *a posteriori* necessity will prove that conceivability does not entail possibility.

The orthodox view on *a posteriori* necessity has been challenged by Chalmers (1996, 2010) and Jackson (1998a). This challenge relies upon a theory of the meaning of referring expressions, known as two-dimensional semantics. According to two-dimensional semantics, referring expressions have two types of content: a primary intension, which normally consists of a descriptive function from possible worlds to objects in them; and a secondary intension, which consists of the object that in fact satisfies the primary intension in the actual world. Two-dimensional semantics allows us to give a deflationary account of the Kripkean *a posteriori* necessities. According to this account, the Kripke cases are ambiguous between the primary and secondary meanings: their primary contents are *a posteriori*, and their secondary contents are necessary – but there is no one interpretation that is both necessary and only knowable *a posteriori*.

If this analysis is right, then the Kripke cases do not prove the orthodox view that conceivability does not entail possibility. They are merely *weak a posteriori* necessities – *weak* because they do not present a scenario, S, such that S is conceivably false but necessarily true. The debate therefore turns on whether or not there are *strong a posteriori* necessities. Strong necessities, if there are any, would be both necessary and *a posteriori* on the same interpretation, and would therefore show conclusively that conceivability does not entail metaphysical possibility.

But are there strong necessities? The current philosophical debate resembles a game of metaphysical whack-a-mole: defenders of the Kripkean orthodoxy propose examples of *a posteriori* necessities that are supposedly not vulnerable to the deflationary two-dimensional account; two-dimensionalists attempt to show that they are in fact vulnerable to the deflationary analysis; another example is proposed; and so it goes on. The game is potentially endless, with both sides able to trade moves, but neither side able to secure a decisive victory. So I will set out an argument to settle the debate, by showing that there cannot be any strong *a posteriori* necessities – because they are impossible.

At the heart of the debate is not just a disagreement over whether this-or-that example can be analysed in two dimensional terms, or even whether two-dimensionalism is a good theory of meaning. There is an even more fundamental dispute about the nature of possibility itself.

At face value, we have two distinct concepts of fundamental possibility: there is a metaphysical notion, which relates to possibility *simpliciter*; and there is an epistemic notion, which is related to the conceivability of a scenario. But what is the relationship between these two types of possibility? Are they really two distinct types? Are are they two ways of looking at the same thing?

There are two opposing views of the matter: modal monism (Chalmers 2010, Schroeter 2012, 2.3), and modal dualism (e.g. Edgington 2004). Modal monism, also known as modal rationalism, is the view that metaphysical modality and an idealised notion of epistemic modality are fundamentally one and the same thing. According to this view, the space of possible worlds is defined by what is ideally conceivable. This is not to deny that there is such a thing as metaphysical necessity – it is just to deny that it is distinct from ideal epistemic necessity. Thus every world that is epistemically possible to an ideal conceiver is metaphysically possible, and vice versa. Modal dualism, in contrast, is the view that metaphysical possibility and epistemic possibility are fundamentally different. If modal dualism is correct, then ideal conceivability need not entail metaphysical possibility.

If there are any genuinely strong *a posteriori* necessities, then modal dualism must be true. The current philosophical orthodoxy is strongly in favour of some form of modal dualism. But I will argue that the current orthodoxy is wrong, and that modal rationalism must be true, because modal dualism is deeply incoherent. And so we should conclude, despite the conventional wisdom, that ideal conceivability does entail metaphysical possibility. This conclusion allows us to employ conceivability arguments against physicalism and externalism, and therefore has profound implications for the metaphysics of mind.

Chapter 1: The Problem of Consciousness

Introduction

The human mind is well-adapted to understanding the world around us. But when the gaze of consciousness is turned upon itself, it is confronted with a mystery. Is our subjective experience itself a physical phenomenon? Or is it something over and above the material world? Strong intuitions pull us in both directions. Thus we might well agree with the physicist Richard Feynman that:

> *Everything is made of atoms*. That is the key hypothesis. The most important hypothesis in all of biology, for example, is that *everything that animals do, atoms do. In other words, there is nothing that living things do that cannot be understood from the point of view that they are made of atoms acting according to the laws of physics*. [1994, p4, Author's italics].

If consciousness is part of the world, and if the world is made of atoms...then surely consciousness is just something that is done by atoms?

And yet, at the same time we might also sympathise with the 19th Century naturalist and advocate of Darwin's theory of evolution, T.H. Huxley, when he wrote:

> […] how it is that any thing so remarkable as a state of consciousness comes about as the result of irritating nervous tissue, is just as unaccountable as the appearance of the Djin when Aladdin rubbed his lamp [1869, p178].

Which view is right? Is consciousness merely something that is done by atoms? Or does the mind stand to the brain as the genie stands to the lamp?

It is probably fair to say that the dominant trend in philosophy has been towards the former view, that is, towards some form of physicalism. But many philosophers have put forward anti-physicalist arguments, attempting to show that consciousness is not (or not just) a physical phenomenon. My view is that the anti-physicalists are right. So I shall argue in this thesis.

The aim of this chapter is to define the problem of consciousness more precisely and explore what would mean for physicalism, or indeed anti-physicalism, to be true. In Section 1.1, I will define the problem of consciousness and state some assumptions that will set the terms of the subsequent enquiry. In Section 1.2, I will define physicalism in relation to the view of consciousness that I advocate, which is property dualism. Section 1.3, I will outline one of the main arguments in favour of physicalism, which I reject, although it is beyond the scope of this thesis to consider it in detail. Finally, in Section 1.4, I will outline an argument, which I develop in this thesis, for the view that consciousness does not supervene on the physical realm, and therefore that physicalism is false.

1.1 - Defining the Problem of Consciousness

What exactly do we mean by consciousness? What is this thing, the very existence of which seems so mysterious? There are lots of different phenomena associated with mind in general, and it is important to distinguish the particular phenomenon of consciousness from those other aspects of mind that may be related, and may present interesting or difficult problems in their own right, but are not quite the same thing. To that end I will use the following definition of consciousness, after Nagel in 'What Is It Like to Be a Bat?' (1974): an entity is conscious if and only if there is something it is like to be that entity. As Nagel puts it:

> […] no matter how the form may vary, the fact that an organism has conscious experience *at all* means, basically, that there is something it is like to *be* that organism. There may be further implications about the form of the experience; there may even (though I doubt it) be implications about the behavior of the organism. But fundamentally an organism has conscious mental states if and only if there is something that it is like to *be* that organism- something it is like *for* the organism.
>
> We may call this the subjective character of experience. [1974, p 436, author's italics]

It is this subjective character, the way feels it to be a particular thing or in a particular state, that constitutes the essential feature of consciousness. This is the mysterious phenomenon that we want to explain – and we need to take care to distinguish it from other mental phenomena. Chalmers (1996, 2010) distinguishes between what he terms the *easy problems* and the *hard problem* of consciousness. The easy problems concern such phenomena as the ability to discriminate, categorize, and react to stimuli, the integration of information by a cognitive system, the deliberate control of behaviour, the difference between wakefulness and sleep, and so forth (2010, p4). Many of these problems concern the biological functions related to consciousness in human beings, and as such one might expect them to be straightforward matters of scientific investigation. Hence they are thought to be relatively easy to understand. Some, but not all, of the supposedly easy problems concern a group of phenomena that we might classify under the heading of *mental representation* – the fact that certain mental states mean something. The hard problem, on the other hand, is to explain the subjective character of experience, phenomenal consciousness. I will explain why this problem is so hard later in this chapter.

The division of the problem of consciousness into easy and hard versions does raise a couple of issues. The first is that the inclusion of mental representation among the supposedly easy problems raises the question of whether there is an essential connection between consciousness and mental representation. If there is an essential connection, then mental representation should arguably be included among the hard problems. But is there an essential connection? I will not address this question in detail in this section – except to note that, very plausibly, the answer is yes and no. There is a sense in which mental representation is independent of consciousness, and a sense in which they are essentially related. In the former sense, representation remains in the category of easy problems; but in the latter sense, it becomes much harder to explain.

Thus on the one hand, there seems to be a perfectly legitimate sense in which representation and associated mental functions can occur without the presence of consciousness. There is a perfectly legitimate sense in which a computer or similar machine may be said to represent objects in the world to the extent that it has internal states that track them, monitor them, process information about them, adjust system outputs accordingly, and so on. In this sense, it seems very plausible that we can give a naturalistic, functional account of representation.

But there also seems to be a sense in which the phenomenon of mental representation is related in some way to consciousness. As well as machine representation, which has no essential connection to consciousness, there seems to be the phenomenon of *conscious representation* - which, as its name suggests, seems at face value to have an intrinsically subjective character. There is nothing it is like to be a computer representing, say, a set of numbers or physical objects; but there is something it is like to be me representing the same things. And, at face value at least, the conscious component of my representational state seems essentially connected to its representational content; it is not an accidental addition. At face value, it seems that there can be something it is like to represent the world as being a certain way – and, moreover, that these two phenomena are connected. In some mental states, the phenomena of representation and consciousness seem essentially related. Of course, this is not uncontroversial – but it does *seem* that way.

These observations raise many further questions; I will try to answer some of these later in this thesis. One set of questions concerns whether it possible to give a functional analysis of conscious representation, as opposed to mere machine representation. And, if it is possible, to what extent would this constitute a functional analysis of consciousness *per se*? I will address this issue in Chapter 5; my view is that there is no possible functional analysis of *conscious* representation. Another set of questions concerns the relationship between the conscious aspect and the representative aspect of conscious representation. Are they really necessarily connected, as I have suggested? And if so, which is the more fundamental? In Chapter 6, I will set out some considerations in favour of what is known as the *phenomenal intentionality theory,* which is the view that the representational content of such states is constituted by their conscious character.

But these arguments about the nature of representation are subsequent to the present argument about consciousness; they build on it, rather than contributing to it. So, for present purposes, I will set aside any consideration of representation as such, and will focus on the phenomenon of consciousness as such, irrespective of whether such conscious states are representative or not.

The second issue with the classification of easy and hard problems of consciousness is a terminological one. In my view, the easy problems (whether or not they include representation) are easy only insofar as they do not constitute or otherwise require an explanation of consciousness; and the explanation of consciousness is hard *per se*. So, rather than speaking of easy and hard problems of consciousness, we ought to say that there are easy and hard problems of mind – and that the hard problem is the problem of consciousness. There is no substantive disagreement with Chalmers here, but for reasons of clarity I will not employ the distinction between easy and hard problems of consciousness in this thesis.

The problem of consciousness – the hard problem of mind – is whether or not the subjective character of experience is fundamentally physical. But what does that mean? What do we mean by *physical*? And what does it mean to assert that consciousness is nothing over and above the physical – i.e. that physicalism is true? In order to get a better handle on these issues, I will make two assumptions that will apply throughout this thesis. The first is an assumption of fact; the second is a matter of definition.

The assumption of fact is that consciousness is part of the objective world. The problem of consciousness that I am concerned with only really arises when we try to understand consciousness as part of the objective world. By this I mean that the subjective perspective of consciousness is itself a part of the world that we encounter via the senses. This does not entail that consciousness is *itself* an object of sense-perception; merely that it is part of the same world as those objects that are. The world contains material objects (whatever they might ultimately turn out to be), and it contains consciousness (which may or may not itself be material). This might seem obvious – but in fact it is not uncontroversial that consciousness is best understood as a proper part of the world that we experience as objective, as 'outer'. For example, according to some interpretations of Kantian idealism (e.g. Walker, 1989), consciousness can never be understood as an object in the world because it is itself the boundary of the world. It is simply that perspective for which the world of objects exists, and which transcends the world and therefore the reach of objective concepts. This view of consciousness raises many problems of its own – but not the problem that I want to address here. Though I will not seek to prove it, I will make the working assumption that consciousness is a proper part of the objective, empirically accessible world and should be understood as such.

Once we have made the assumption that consciousness is part of the same objective world as physical objects, the question then arises as to whether or not consciousness itself is fundamentally a physical phenomenon. Thus, in simple terms, the problem of consciousness is whether or not the objective world at its most fundamental level is just physical, or physical plus something else. This of course raises the question of what exactly we mean by *physical*.

This leads to the matter of definition. For the purpose of this thesis, I will define the fundamental physical realm to be whatever entities, properties, laws, and so forth, an idealised complete physics would need to posit to explain the objective world, where physics is defined as the most fundamental empirical science of the world. I will not speculate on what these entities or laws may turn out to be. Perhaps an ultimate, unified physics would look very different from any theory we have today; it does not matter. There is however one important caveat: I will define an idealised physics to include, at the fundamental level, only non-mental entities, properties and events, and the impersonal laws of their combination (cf Papineau 2002, p41). What do I mean by mental? In this context, I take an entity, event or property to be mental if and only if it is fundamentally conscious or confers consciousness on its holder.[1]

It is important to note that the physical is defined as that which is non-mental at the fundamental level – and not as the non-mental *simpliciter*. The latter definition would prejudice the whole enquiry from the outset. If we define the physical as non-mental *simpliciter*, then how can consciousness be a physical phenomenon? But this would be too quick. The question is whether consciousness is ultimately constituted by more basic, non-mental items. There are many complex phenomena in the world that are not present at the level of fundamental physics (so far as we know), but which do seem to be constituted by nothing more than physics. For example, biological life is constituted by non-living fundamental physical processes. We now know that life is not fundamental to the universe; it is made up of non-living stuff. So the real

[1] This way of defining the physical solves what has become known as *Hempel's Dilemma* (Hempel 1980, pp194-195). There are, according to the dilemma, two ways to define the fundamental physical realm: either physical properties are defined in terms of present physics, or they are defined in terms of an idealised, completed physics. If we take the first option, then it is highly likely that our understanding of what physical properties are like will ultimately prove false, and that it will therefore be false that consciousness can be explained in terms of material objects thus (wrongly) defined. But if we take the second option, then we cannot know in advance what the completed physics will look like – perhaps it will attribute to nature properties that we would not recognise as physical at all, in which case we do not even know what it amounts to say that consciousness can be reduced to matter. The dilemma is solved by taking the second horn of the dilemma, but adding, as I have done, that an idealised physics does not include any fundamentally mental or proto-mental properties. It does not matter what the fundamental physical realm turns out to be like, as long as it is not mental. The question is then whether the physical, thus defined, can be sufficient for consciousness.

It might be objected that we could avoid Hempel's dilemma entirely by simply taking the *via negativa* and defining the physical as the non-mental, without any reference to physics, present or ideal. But I think this will not work, because there could be non-mental items that have nothing to do with the constitution of the objective world, such as epiphenomenal non-mental goo, or abstract objects, and we would not want them to count as physical. So we first need to restrict the domain to that part of reality that is investigated by physics – the stuff, whatever it may be, that makes up tables, chairs, and human animals – and only then do we apply the *via negativa*.

question is whether consciousness is like life. Is it made up of non-mental stuff? Or is it – unlike biological life – somehow fundamental to reality?

I will argue in this thesis that physicalism is false. But what are the alternatives? If it turns out that consciousness is not reducible to the physical realm, as I have defined it, then we must conclude that it is itself a fundamental component of reality.

There are two broad ways in which this might be true. The first is some form of dualism: in addition to the fundamental physical (i.e. non-mental) realm, we must posit some fundamental conscious entities or consciousness-related principles. The second is panpsychism, which is the thesis that the objective world – the world investigated by physics – is fundamentally conscious or proto-conscious (Goff, Seager, and Allen-Hermanson, 2017)

If panpsychism is true, then the objective world – the world investigated by physics – will not count as fundamentally physical by my definition. Panpsychism draws a distinction between the causal roles that constitute the observable physical world – such as the *charge* role, the *mass* role, and so forth – and the fundamental entities (whatever they may be) that play these roles. The laws of physics describe the roles; but they in turn are ultimately grounded in the intrinsic properties of the role-players. The idea is then that the intrinsic properties of the role players are such that, as well as grounding the laws of physics, they are themselves conscious.

The advocates of panpsychism, such as Coleman & Alter (2018), claim that it avoids the problems facing both standard physicalism (which I will set out in Section 1.4 and argue in this thesis) and conventional dualism (as I shall outline in Section 1.3). But panpsychism has its own problems. The most significant is the combination problem: even if it is granted that the causal roles of physics are grounded in fundamental consciousness, and therefore that consciousness is everywhere at the fundamental level, this does not explain why higher-order consciousness is found in some large-scale arrangements of matter and not others. In simple terms, the microphysical particles that constitute a plant are just as much grounded in consciousness as those that constitute a living human body; yet a human body is (normally) a host to a higher-order, unified consciousness – and plants, so far as we know, are not. Why do some arrangements of matter lead to the emergence of higher order consciousness, but not others?

Perhaps it is possible for panpsychism to solve the combination problem. I take the view that panpsychism is a real metaphysical possibility and a serious attempt to answer a genuine question, about the fundamental nature of objective reality. But for the purpose of this thesis I will assume that panpsychism is false, and that the causal roles of the fundamental physical realm are not grounded in consciousness.

Similarly, I will set aside *panprotopsychism*. This is the view that the intrinsic nature of the fundamental role-players, whilst not in itself conscious, is such that it *a priori* entails the emergence of higher-order consciousness in the appropriate circumstances. How does panprotopsychism differ from physicalism? The normal understanding of physicalism is that the fundamental causal roles themselves, as opposed to some hypothesised intrinsic role-players, are metaphysically sufficient for consciousness when combined appropriately. This is what I mean by physicalism in this thesis, and this is the hypothesis that I will examine and ultimately reject. As with panpsychism, I consider panprotopsychism to be metaphysically possible, and I am neutral as to whether it is true or not.[2] But, for the purpose of investigating the form of version of physicalism that is my intended target in this thesis, I will not consider panprotopsychism.

Having set aside panpsychism and panprotopsychism aside for the purpose of this thesis, it follows that if physicalism as I have defined it is false, then dualism is true. So, to outline the shape of problem of consciousness, we have the following points:

a) The defining characteristic of consciousness is that there is something it is like to be conscious.

b) Consciousness may or may not be closely related to the phenomenon of mental representation, but I will postpone discussion of these issues until later in the thesis, and focus for now on the problem of explaining consciousness *per se*.

[2] My view, as I will explain later, is that there is an *a priori* entailment from fundamental physics plus contingent psycho-physical connecting laws to consciousness; I regard this view as a form of naturalistic dualism. I tend to view the psycho-physical connecting laws as existing alongside or on top of the laws of physics. But I see no reason in principle why they could not be located 'inside' the fundamental physical role-players or grounded in their intrinsic nature – which would arguably be a form of panprotopsychism.

c) I will assume for the sake of argument that consciousness is a proper part of the world that we experience as external in sense-perception (which is not the same thing as saying that consciousness is itself an object of sense-perception), and so exclude from consideration any form of transcendental idealism or similar.

d) The physical realm is defined as being the world as revealed by an idealised final empirical science, as long as this idealised physics does not posit any fundamental conscious entities.

e) The problem of consciousness is therefore whether the facts of the actual physical realm, thus defined, are sufficient for consciousness.

f) If the physical facts, thus defined, are sufficient for consciousness, then physicalism is true and consciousness is ultimately just a physical phenomenon; if not, then either dualism or panpsychism or panprotopsychism is true.

g) For the purposes of this thesis, I will set aside panpsychism and panprotopsychism. The question is therefore whether physicalism or dualism is true.

1.2 – Physicalism v Dualism

Physicalism is the thesis that the physical properties of the actual world are metaphysically sufficient to determine the facts about consciousness in the actual world. Intuitively, we might imagine God creating the physical facts of the actual world. If physicalism is true, then this is *all* God had to do in order to make it the case that the facts about consciousness were just as they actually are. Once the physical facts are fixed, the facts about consciousness follow automatically, as it were. They are nothing 'over and above' the physical facts. Following Jackson (1998a), this can be expressed in terms of possible worlds: physicalism is true (in the actual world) if and only if there is no possible world that is a minimal duplicate of the actual world with respect to its physical properties, but which fails to duplicate the facts about consciousness. It is important to note from this definition that even if physicalism is true in the actual world, this does *not* entail that there are no worlds in which non-physical consciousness exists. Physicalism in the actual world is consistent with the existence of other possible worlds containing immaterial consciousness. The point of physicalism is not that the physics of the actual world is metaphysically *necessary* for consciousness, but that it is metaphysically *sufficient*.

Physicalism stands in contrast with dualism. Dualism, at its most basic, is the claim that the physical world on its own is not sufficient for consciousness, and therefore that mental consciousness also requires the existence of some non-physical fundamental reality. There are different types of dualism, involving different strengths of commitment. The strongest version is substance dualism. This is the view that conscious mind and physical matter are separate and distinct types of fundamental substance, which exist alongside one another in the inventory of the fundamental building blocks of reality. But I will not advocate substance dualism in this thesis. My aim is only to show that the minimal form of dualism that stands opposed to physicalism must be true – and the minimal form of dualism is property dualism. This is the thesis that the property of being conscious, or of having a particular conscious state, is not a physical property (even a very complex one), nor is it something that is constituted or metaphysically determined just by physical properties. If substance dualism is true, then, *a fortiori*, property dualism must be true. But property dualism (at least at face value) does not entail substance dualism: it apparently leaves open the possibility that, under the right conditions, physical substances (such as living human bodies) can instantiate non-physical mental properties. It seems *prima facie* possible to combine property dualism with the denial of mental substances. Perhaps all substances are physical, but some physical substances happen also to have mental properties, which are not metaphysically determined by the distribution of physical properties in the world. But I will not dwell on the issue here. In this thesis I will argue only for the minimal dualist claim, which is property dualism.

Both physicalism and property dualism can be understood in terms of the notion of *supervenience*. Lewis (1986) explains supervenience by analogy with an image produced by a dot-matrix printer:

> A dot-matrix picture has global properties — it is symmetrical, it is cluttered, and whatnot — and yet all there is to the picture is dots and non-dots at each point of the matrix. The global properties are nothing but patterns in the dots. They supervene: no two pictures could differ in their global properties without differing, somewhere, in whether there is or there isn't a dot. [p14]

The distribution of dots corresponds to the distribution of microphysical properties. The global properties correspond to the macrophysical properties. Any two pictures that are alike in respect of the distribution of dots must necessarily be alike in respect of their symmetry and so forth. Thus there is a sense in which the global properties are nothing over and above the underlying distribution of dots.

There are different types of supervenience. We can distinguish between *global* and *local* supervenience relations. Global supervenience claims assert that the totality of the lower-order facts in the world are sufficient for the totality of the higher-order facts; local supervenience claims assert that some portion of the lower-order facts are sufficient for a corresponding portion of the higher-order facts. In the case of physicalism, global supervenience would entail that the totality of the physical facts in the world are sufficient for the sum total of facts about consciousness; local supervenience, on the other hand, would mean that some local portion of the physical facts – for example, the facts about my brain states – are sufficient for some corresponding portion of facts about consciousness, namely my consciousness. This distinction is relatively unimportant for present purposes. It is natural to regard physicalism as implying a local form of supervenience of consciousness (though not necessarily mental content, as I will discuss in Chapters 6 and 7) on the physical realm – that my brain should be sufficient for my consciousness – and also as making a claim about the world as a whole. Conversely, the argument against physicalism – which I will outline in Section 1.4 – works against physicalism in both its local and global forms. So I will not make further use of this distinction in this thesis.

Of much greater importance is the distinction between strong, or metaphysical, and weak, or causal supervenience (cf McLaughlin and Bennett 2014, 3.1). Metaphysical supervenience means that the distribution of physical properties in the actual world is sufficient to fix the distribution of mental properties as a matter of metaphysical necessity. That is, in any possible world in which the distribution of physical properties is identical to the actual world, there must also be an identical distribution of mental properties. Conversely, if there is a possible world in which all the physical properties are exactly as they are in the actual world, but which does not have any mental properties, or in which the mental properties are different, then metaphysical supervenience is false. Causal supervenience, on the other hand, does not imply any such metaphysical entailment. It is a claim about the actual world, rather than the space of possible worlds: causal supervenience is the thesis that the physical properties of the actual world are causally sufficient for consciousness, perhaps via contingent laws of nature.

We can define physicalism in terms of metaphysical supervenience: it is the thesis that consciousness metaphysically supervenes on the fundamental physical properties of the world. Therefore, according to the formulation of Jackson (1998a) physicalism is true if and only if there is no possible world that is a (minimal) physical duplicate of the actual world, but which differs with respect to consciousness. This formulation will play an important role in the central argument against physicalism, which I will outline in Section 1.4 and consider in detail in this thesis. If physicalism, thus defined, is false, then, putting aside the possibility of panpsychism, some form of dualism will be true.[3]

Although property dualism is incompatible with the metaphysical supervenience of consciousness on the physical, it is perfectly consistent with causal supervenience. Dualism is compatible with the existence of causal laws that connect mental and physical events – and therefore it is perfectly consistent with a naturalistic account of mind, such as an evolutionary explanation of the origins of consciousness. The only constraint on such a naturalistic account is that it cannot be a purely physical account. We might, for example, give a purely physical explanation of the evolution of certain brain functions – but, if dualism is true, then we would also need to invoke psycho-physical connecting laws to explain the contingent connection between these brain functions and consciousness.

So, we can summarise the last two sections, and the relationship between property dualism, physicalism and supervenience as follows:

[3] It is important to note what physicalism, thus defined, does *not* entail: it does not entail that the physical facts of the actual world are metaphysically *necessary* for consciousness. Physicalism is just the view that consciousness as it exists in the actual world is nothing metaphysically over and above the physical facts of the actual world; this is more formally stated in terms of the physical facts being metaphysically sufficient for consciousness. But it is a further claim that conscious properties are identical to their actual physical realisers. As long as physicalism is not committed to this additional, stronger claim, it leaves open the possibility that consciousness could (metaphysically) exist without physical matter (perhaps there are worlds of pure spirit that are exact duplicates of the actual world with respect to consciousness). Similarly, we are all physicalists about the watery stuff in our world – it's just H_2O. The existence of H_2O is metaphysically sufficient for the existence of watery stuff. But this does not rule out the possibility of worlds containing watery stuff that is not H_2O.

a) Physicalism is true of some realm of higher-order facts (in this case, the facts about consciousness) just if these higher-order facts supervene metaphysically on the microphysical facts.

b) Property dualism is true of some realm of higher-order facts just if they do not supervene metaphysically on the microphysical facts.

c) Property dualism is compatible with the causal supervenience of the higher-order facts (in this case, facts about consciousness) on microphysics.

We now have a clear picture of what physicalism is, and what its truth or falsehood would mean for the actual world. But is it true?

1.3 - The Causal Argument For Physicalism

It is beyond the scope of this thesis to consider every argument for and against physicalism in any great detail. So I will focus the central argument against physicalism, which I will outline in Section 1.4 and consider in detail in subsequent chapters. In the present section, I will outline the central argument in favour of physicalism – namely the causal argument, also known as the epiphenomenalism problem. Although it is not my intention to explore it in detail – still less offer a solution – it is worth discussing for the sake of context.

The causal argument for physicalism arises from the fact that it seems highly likely, as a matter of empirical fact, that the physical world is causally closed. What does this mean? The principle of causal closure states that every physical event has a sufficient cause that is also physical.[4] If

[4] The principle of causal closure, thus defined, allows for the possibility of over-determination of a physical event by both a physical and a non-physical cause, each of which would be individually sufficient for the effect. Thus this principle might be better termed the causal *completeness* of the physical – because, while causally complete, the physical realm is not in fact completely closed. However, while I think the term *completeness* better describes this principle, I will stick with the terminology that is more current in the literature. A stronger version of causal closure – and one which would truly merit the name – states that no physical event has a non-physical cause. This stronger version rules out over-determination by non-physical causes and, when combined with dualism, leads directly to the epiphenomenalism of consciousness.

every physical event has a sufficient cause that is itself physical, this seems to render non-physical causes superfluous with respect to physical effects. But if non-physical causes are superfluous, then we seem to have the following dilemma: either they are causally irrelevant to physical events – a position known as *epiphenomenalism* – or some physical events are over-determined, in that they have both physical and (superfluous) non-physical causes. So non-physicalism about consciousness, plus the causal closure of the physical, seems to entail that either consciousness is epiphenomenal, or that some physical events are over-determined.

So what? One might wonder what is wrong with epiphenomenalism and overdetermination. There are various reasons, which I will not go into here, why both epiphenomenalism and over-determination are often regarded as being implausible or undesirable. Perhaps it is not so easy to say exactly why they both must be false; but, for the sake of argument, I will assume that they are both consequences that an adequate account of consciousness would need to avoid. But this assumption, when combined with the principle of causal closure of the physical, suggests an argument for physicalism along the following lines: both epiphenomenalism and over-determination are false; therefore causal closure of the physical and dualism cannot both be true; causal closure of the physical is true; therefore dualism is false; therefore (if we discount panpsychism) physicalism is true.

Papineau (2002, pp17-18) sets out the argument in the following terms:

> (1) Conscious mental occurrences have physical effects.
>
> […]
>
> (2) All physical effects are fully caused by purely *physical* prior histories. [Author's italics]

Papineau adds a further premise to rule out the possibility that physical events are causally over-determined, with both a physical cause and a non-physical (i.e. conscious mental) cause:

> (3) The physical effects of conscious causes aren't always over-determined by distinct causes.

From these premises, he argues, materialism follows. Conscious mental occurrences have physical effects; physical effects are caused only by prior physical causes; therefore conscious mental events are physical. So the causal closure of the physical seems to offer a powerful argument for physicalism. At the very least, these considerations seem to present an unattractive set of options for dualism: either accept the epiphenomenalism of consciousness; or accept the over-determination of some physical effects by both mental and physical causes; or reject the causal closure of the physical.

However, the picture is rather more complicated. It is not clear that the problems associated with causal closure are limited to dualist accounts of consciousness. Kim (1989) argues that the causal argument poses exactly the same problem for non-reductive physicalism (that is, physicalism which allows for multiple instantiation of mental properties, such as minimal supervenience physicalism) as it does for dualism. If Kim is right, then non-reductive physicalism, like dualism, leads to a forced choice between epiphenomenalism, over-determination, and rejecting causal closure. Why does he think this? The idea is that if mental properties are not *identical* to physical properties, then we cannot allow them causal relevance to physical events without either giving up on causal closure or allowing over-determination. Merely allowing that mental types supervene metaphysically on physical types, as non-reductive physicalism does, will not be enough to avoid this problem. As Kim puts it:

> Given that any physical event has a physical cause, how is a mental cause *also* possible? This I call "the problem of causal-explanatory exclusion", for the problem seems to arise from the fact that a cause […] of an event […] appears to *exclude* other *independent* purported causes or causal explanations of it.
>
> […] Why not just say the mental cause and the physical cause are one and the same? Identification simplifies ontology and gets rid of unwanted puzzles. […] what is at issue is the causal efficacy of *mental properties* of events vis-a-vis their physical properties. […] we would need to identify mental properties with physical properties. If this could be done, that would be an excellent way of vindicating the causal powers of mentality.
>
> But this is precisely the route that is barred to our non-reductivist friends. [1989, pp44-45, author's italics]

Kim intends this argument to favour reductive physicalism (such as type-identity physicalism) over non-reductive versions – after all, if non-reductive physicalism leads to the same problems as dualism, then we should reject it for the same reasons.

But there is a further problem, which is that Kim's argument seems *too* strong: it generalises to any macroscopic properties that are capable of instantiation by different physical realisers. The same considerations that apply to consciousness if non-reductive physicalism is true will apply equally to any realm of facts that is functionally defined and multiply realisable, but whose types cannot be identified with physical types. But this threatens many macroscopic properties – including mechanical biological, meteorological properties and so forth, none of which are plausibly identical to any physical types – with causal irrelevance on pain of over-determination or violating the causal closure of the physical (cf Robb & Heil, 2019).

What is even worse (from a physicalist point of view) is that Kim's problem apparently extends to reductive physicalism as well. How so? Surely if mental properties or states are identical to physical ones, then they should not be affected by the exclusion problem? But the problem is that reductive physicalism does not generally posit identities between mental states and microphysical states, but between mental states and macroscopic physical states that are themselves affected by Kim's version of the exclusion problem. *A fortiori*, the mental states themselves are excluded. For example, suppose that pain is identical to C-fibre stimulation. Now the problem is that C-fibre stimulation itself is multiply realisable by different underlying microphysical structures, and the macro-property of being C-fibre stimulation seems to be excluded from causal relevance by the closure of the underlying microphysical causes.

Does this mean that the exclusion problem is not unique to dualism? If it really does generalise to physicalism, as Kim's argument suggests, then this would make the causal argument for physicalism redundant. Unfortunately (from the point of view of dualism), I do not think this is the case. The likelihood is that it is possible to resist Kim's argument, and therefore that there is a distinctive exclusion problem for dualism. For example, Bennett (2008) argues that certain kinds of tightly related causes can both be causally sufficient for the same effect without overdetermining it – and that this applies if some form of physicalism is true (including the non-reductive version that she favours), but not in the case of dualism.

I think an approach along these lines is likely to be successful. In this case, dualism will face a distinctive exclusion problem, and the classic causal argument for physicalism will present a real dilemma for dualism: either epiphenomenalism is true, or causal closure is false. As to which horn of this dilemma I, as a dualist, think we should embrace – I will leave that an open question for now.

1.4 - The Zombie Argument Against Physicalism

Now I will turn to what I consider to be the central argument against physicalism. Chalmers's two-dimensional argument against physicalism (Chalmers 2010, pp141-205) – also known as the zombie argument – aims to show that consciousness does not metaphysically supervene on the physical properties of the actual world, and therefore that property dualism is true. It aims to do this by demonstrating that there is a metaphysically possible world that is an exact duplicate of the actual world with respect to all of the physical properties, but which does not contain consciousness.

The argument begins by inviting us to imagine a possible world inhabited by creatures known as *philosophical zombies.* Philosophical zombies are not like the creatures from the movies and TV shows – on the contrary, a *philosophical* zombie is outwardly completely indistinguishable from an ordinary human being. This is because an ordinary human and their zombie counterpart are qualitatively identical with respect to all of their physical properties. In fact, they are imagined to be atom-for-atom replicas (or, to be more precise, they are qualitatively identical with respect to their fundamental physical properties, whatever these may ultimately turn out to be). Thus a zombie in this sense is a minimal physical duplicate of a human being (though not necessarily a complete *functional* duplicate – that of course will depend on whether or not all human functions are purely physical functions). But – and this is the crucial difference – there is nothing it is like to be a philosophical zombie. These zombies have exactly the same brain processes and so forth as ordinary humans, but they have no subjectivity. There is no first-person viewpoint associated with these beings.

The argument does not require zombies to be causally possible in the actual world. Indeed, if we make certain assumptions about the uniformity of nature and its laws, then it seems plausible that anything that had human brain processes would also have human consciousness. The two-dimensional argument does not rule out the causal or natural supervenience of consciousness on physical properties (cf 2010, pp 126-130); but it does entail that supervenience can *only* be natural or causal – that is, it is a function of the contingent laws that exist in the actual world, and not a feature of all possible worlds.

The argument thus moves from a premise about the conceivability of zombies to a claim about the metaphysical possibility of a zombie world. The next step moves from the possibility of a zombie world to the falsehood of physicalism. So we can summarise the overall structure of the argument as follows (cf 2010, p142):

i) If zombies are metaphysically possible, physicalism is false.

ii) Zombies are ideally conceivable.

iii) Ideal conceivability entails metaphysical possibility.

Now, (ii) together with (iii) entails:

iv) Zombies are metaphysically possible.

And (i) together with (iv) entails:

v) Physicalism is false.

Does the argument succeed? This, of course, depends on its validity, and on the truth or otherwise of its premises. At face value, the argument seems valid. Nevertheless, this has been questioned, and I will address objections to the argument's validity in Chapter 5. What about the premises? A physicalist opponent of the argument might object to any of premises (i), (ii), or (iii). I will briefly address these in reverse order. By far the most important objection to the argument is that ideal conceivability does not entail metaphysical possibility – that premise (iii) is false. This is associated with what Chalmers calls *Type B* physicalism, which is the view that zombies are conceivable, but not metaphysically possible. Type B physicalism, in other words, holds that there is a metaphysical entailment from the physical facts to consciousness, even though there is no *a priori* entailment (2010, p 115). Strictly speaking, one can reject the entailment from conceivability to metaphysical possibility without being a Type B physicalist (one could have other grounds for thinking that zombies are metaphysically possible); but one cannot be a Type B physicalist without rejecting the entailment. I will address this objection at length in Chapters 2 to 4.

The second most important objection is that premise (ii) is false – that zombies are not really conceivable. This leads to what Chalmers terms *Type A* physicalism: the view that there is an *a priori* entailment from the physical facts to consciousness (2010, p 111). I will address this objection in Chapter 5. Finally, one could theoretically oppose the argument by rejecting premise (i). But this is a less serious objection and can be dealt with relatively easily – so I will address it now.

Why would the metaphysical possibility of zombies entail that physicalism is false? This follows from the definition of physicalism in terms of supervenience. If zombies are possible then consciousness does not supervene metaphysically on the physical properties of the world. And, as I argued above, if consciousness does not metaphysically supervene on the physical properties of the actual world, then property dualism is true. It is for exactly this reason that the standard formulation of physicalism defines it in terms of metaphysical supervenience (cf Jackson 1998a). Metaphysical supervenience means that any possible world that is a minimal physical duplicate of our own world is a duplicate *simpliciter* – and therefore that it is a duplicate with respect to consciousness.

We can illustrate this as follows: let *P* represent the total distribution of physical properties in the actual world, and *Q* represent the facts about consciousness as they are in the actual world (notation is as per Chalmers 2010, pp141-205).

Thus in the actual world, we have:

$$P \& Q$$

But in the zombie possible world the following is true:

$$P \& {\sim}Q$$

But the supervenience of Q on P means that P is sufficient to determine that Q. At a first attempt, we might express the supervenience relation thus:

$$P \Rightarrow Q$$

But this is not quite right. The problem is that we have to allow for the possibility of what in the literature are known as *blockers* (cf Hawthorne 2002). A blocker is a physical fact which, when added to the supervenience base P, prevents the emergence of mental facts Q. Thus we might imagine a possible world that is an exact physical replica of our own, except that it also contains, in addition to the total distribution of matter and energy found in our own world, one magical atom on the far side of the universe, which has the effect of suppressing the occurrence of consciousness on Earth. Such a universe, whilst strange, does appear to be conceivable. But it is a universe in which P is true, but Q is not. Hence it appears to show that P is not sufficient for Q. At first glance, this might appear to be an example of a zombie-universe. But this is misleading. The possibility of a blocker universe does not in itself undermine physicalism, and it does not entail the possibility of a zombie universe. Even if blockers are possible, it may still be the case that the distribution of physical properties found in *our* world is metaphysically sufficient to fix the mental properties as they occur in our universe – because our universe happens not to contain blockers. We can remove the problem of blockers up by the addition of what Chalmers calls a 'that's all' clause (2010, p143), *T* – which is, in effect, a stipulation that the supervenience base does not contain blockers. So a more accurate statement of the supervenience relation would be:

$$PT \Rightarrow Q$$

In order for physicalism to be true, the arrow from PT to Q has to represent metaphysical necessitation. That is, there must be no possible world in which PT obtains, but not Q.

A similar complication to the blockers problem arises from the apparent possibility of *epiphenomenal ectoplasm* (cf Jackson 1998a, Stoljar 2009). This refers to the existence of consciousness in a possible world which has no connection, causal or metaphysical, to the physics of that world. It is there, but is completely inert and completely independent, at least in relation to physics. Now, it seems at face value that there is a possible world that is a minimal physical duplicate of the actual world, with no blockers, that has all the same mental properties of the actual world – plus epiphenomenal ectoplasm. Why is this a problem? The problem arises because physicalism, as it is typically formulated, seems to rule this out: physicalism is true if and only if there is no possible world that is a minimal physical duplicate of the actual world (containing no blockers), but which differs with respect to consciousness. But the epiphenomenal ectoplasm world, although it is a minimal physical duplicate of the actual world, with no blockers, does differ from the actual world with respect to consciousness: it contains epiphenomenal ectoplasm. So it seems that the ectoplasm world, like the zombie world, offers a refutation of physicalism. But, unlike the zombie world, it seems intuitively that it ought not to. The whole point of physicalism is that the physics of the actual world is metaphysically sufficient for the consciousness found in the actual world, not that it metaphysically rules out additional, epiphenomenal consciousness.

There are several ways to respond to the epiphenomenal ectoplasm problem. One way is to take the specification that a world be a *minimal* physical duplicate of the actual world to rule out, not just blockers, but epiphenomenal ectoplasm as well. An alternative approach, offered by Chalmers (2010) is to specify only that, if physicalism is true, then a minimal physical duplicate of the actual world must also be a duplicate with respect to its *positive* conscious properties. That is, such a world must not lack any consciousness found in the actual world – but whether or not it also contains ectoplasm is a further, irrelevant consideration. Along these lines, we might refine the definition of physicalism as follows: physicalism is true if and only if there is no possible world that is a minimal duplicate of the actual world, but which fails to duplicate the consciousness found in the actual world. Whichever approach is preferred, I do not think the problem is a substantive one. It is simply a matter of formulating, or interpreting, the definition of physicalism in such a way that it rules out zombies but not ectoplasm.

In summary, we can understand these issues in terms of the following parable (cf Chalmers 2010, also Kripke 1980): if metaphysical supervenience is true, then all God had to do in order to create a world in which consciousness exists is to create a world that is physically qualitatively identical to the actual world, and which does not contain blockers. Given these conditions, the existence of consciousness necessarily follows. Since there is no possible world in which these conditions are met but there is no consciousness, there was no more work for God to do. In this case, physicalism is true. But if metaphysical supervenience is false, then it would not be enough for God to create a universe with identical physical properties to the actual world, and no blockers. This would leave open the possibility that the resulting world would be a zombie world. Hence there would be more work for God to do in order to bring about consciousness – perhaps by creating psycho-physical connecting laws, which we represent as *L*. These are contingent laws that, when combined with PT, are sufficient for Q. Such psycho-physical connecting laws are by definition not part of the physics of the actual world – they are an addition to P, not contained within it. So physicalism would still be false in this case.

It is important to bear in mind the distinction between metaphysical supervenience, which is necessary and sufficient for physicalism to be true, and mere causal supervenience, which is consistent with dualism. Just as causal supervenience is not sufficient for physicalism, so the denial of physicalism does not entail that there are no law-like connections between consciousness and the physical world. We can represent causal supervenience as follows:

$$P + L + T \Rightarrow Q$$

If physicalism is false, then it will be a contingent matter whether or not there are psycho-physical connecting laws in the actual world, to be determined by empirical enquiry.

So premise (i) of the zombie argument is undoubtedly true. But, as I set out above, this still leaves three possible ways for a physicalist to object to the argument and resist the conclusion: first, to show that it is somehow invalid; second, to deny that zombies are in fact ideally conceivable; and third, to deny that ideal conceivability entails metaphysical possibility. I will address the first two possible objections in Chapter 5. But in the next three chapters I will look in detail at the most important and controversial issue of all, which is whether ideal conceivability entails metaphysical possibility.

Conclusion

The essential characteristic of consciousness is that there is something it is like to be conscious. This is the mysterious phenomenon that we are trying to understand. For the purpose of this thesis, I make several assumptions about the place of consciousness in the world, and the nature of the world itself. These assumptions serve to frame the problem. The first is that consciousness is a part of the objective, external world that we encounter in sense-perception – that it is a feature of the ordinary world, existing alongside everyday physical objects. In other words, I assume that idealism is not true. But this raises the question: what is the physical world itself? So I also make the assumption that the physical world is whatever an idealised physics reveals it to be, so long as that idealised physics does not make any reference to mental or proto-mental items.

Given these assumptions, the question is whether consciousness is an ultimately physical phenomenon, analogous to biological life – in which case physicalism will be true. The alternative is that consciousness is a fundamental constituent of reality, existing alongside the physical realm – in which case some form of dualism will be true. The minimal form of dualism is property dualism, which is the claim that the totality of physical properties of the universe are not metaphysically sufficient to determine its conscious properties. Property dualism is consistent with law-like correlations between physical properties and consciousness; and it does not entail substance dualism, being consistent with the proposition that conscious properties are instantiated by physical substances, such as living bodies.

I accept the standard formulation of physicalism in terms of the metaphysical supervenience of consciousness on the physical properties of the world. Therefore physicalism can be expressed as a claim about the space of possible worlds: that there is no possible world that is a minimal physical duplicate of the actual world, but which fails to duplicate the consciousness found in the actual world. The zombie (or two-dimensional) argument against physicalism claims to show that there is such a possible world, and therefore that physicalism is false. In its simplest form, the argument is as follows: if physicalism is true then there is no possible zombie world; it is possible to conceive of a zombie world; conceivability entails possibility; therefore there is such a possible world; therefore physicalism is false.

There are several grounds on which one might object to the zombie argument. One might claim that the argument is not valid, or that zombies are not really conceivable. I will address these objections (and others) in Chapter 5. But the most important objection to the zombie argument – indeed, to conceivability arguments in general – is that conceivability does not entail possibility. That is the issue that I will address in the next three chapters, where my aim is to show that, contrary to the conventional wisdom, ideal conceivability does entail metaphysical possibility.

Chapter 2: Does Conceivability Entail Possibility?

Introduction

Does the conceivability of a scenario entail its metaphysical possibility? Since the work of Kripke (1971,1980), the philosophical orthodoxy has been that it does not. It is now commonly accepted that there are *a posteriori* necessities – that is, that there are some propositions that are true in all possible worlds, but whose truth can only be known *a posteriori*, even by an ideal rational subject. The negation of such a truth would be conceivably true, but necessarily false. And therefore, so the argument goes, conceivability does not entail possibility.

But this view, though widespread, is not universal. Chalmers (2010) (also Jackson, 1998b) argues that this view is the result of misunderstanding the nature of the referring expressions involved. His account involves a theory of meaning and reference known as *two-dimensional semantics*. According to this account, the referring expressions involved, and thus the supposed *a posteriori* necessities, are ambiguous. The Kripke cases have one sense in which they are necessary; and another sense in which they are *a posteriori;* but no sense in which they are both necessary and *a posteriori*. Therefore the Kripke cases do not really disprove the entailment from conceivability to metaphysical possibility.

I agree with this analysis, and will argue for a further, stronger claim: there *cannot* be any unambiguous *a posteriori* necessities. They are impossible – in fact, they are inconceivable. The central claim of this thesis is that the idea of unambiguous *a posteriori* necessities is incoherent, and therefore that an appropriate notion of conceivability does indeed entail metaphysical possibility. The current philosophical orthodoxy about the fundamental nature of modality is wrong.

I will make that argument in Chapter 4. The aim of this chapter is to set out the issue, and show that the Kripke cases are not really decisive. I will prove a double negative: that the Kripke cases do not prove that conceivability does not entail metaphysical possibility. In Section 2.1, I will set out various notions of conceivability and possibility, in order to define the problem more precisely. In Section 2.2, I will explain the Kripkean *a posteriori* necessities and why these are widely thought to show that conceivability does not entail possibility. In Sections 2.3, I will set out the two-dimensional semantic framework, and in Section 2.4, I will outline the deflationary, two-dimensional analysis of the Kripke cases.

2.1 - Conceivability and Possibility

At first glance, it might seem absurd raise the question of whether conceivability entails possibility. Of course it does not. It is easy to conceive of things that are impossible in one way or another. However, I will argue that there is an important sense in which conceivability does entail possibility – and that substantial philosophical consequences follow from this. So first we need to be clear on exactly what it is that I am claiming, and what my opponents deny. In order to be clear on this, we need to define the notions of conceivability and possibility more precisely.

With respect to conceivability, Chalmers (2010, pp143-147) draws a distinction between *prima facie* conceivability – that is, what seems, on first appearances, to be conceivable to a subject such as myself, with limited intellectual powers – and *ideal* conceivability – that is, what would be conceivable to an idealised rational thinker who has no cognitive limitations or imperfections.

How should we understand the distinction between these two notions of conceivability? We should not think of ideal and *prima facie* conceivability as merely two types of conceivability, of which one happens to be superior to the other. This is a misinterpretation, from which much confusion is liable to follow. On the contrary, ideal conceivability is better characterised as conceivability *simpliciter*. If a scenario is ideally conceivable, this just means that there is a coherent conception of the scenario in question. This just means that there is no logical contradiction in what is conceived or proposed, and it does not entail a logical contradiction. Truths about ideal conceivability are ultimately grounded in truths of logic. Mere *prima facie* conceivability, on the other hand, is not an inferior type of conceivability, so much as the *illusion* of conceivability. If something is merely *prima facie* (so not ideally) conceivable, this means that there is not really a coherent scenario in question at all – in other words, there is a logical contradiction contained within or entailed by the *prima facie* conception. There is only the appearance of one. So the difference between ideal and mere *prima facie* conceivability is not the difference between two distinct ways of conceiving, one better than the other. Rather, it is the difference between genuine conceivability and the mere illusion of conceivability. Accordingly, an ideally conceivable scenario is a scenario *simpliciter*; a mere *prima facie* conceivable scenario is no scenario at all, but the illusion of one.

Although we are not ourselves ideal conceivers, we are sometimes (though not always) able to draw reasonable conclusions about what is or is not ideally conceivable. Thus we can be certain that it is not ideally conceivable that the square root of 2 is rational, but that (for example) a flying horse or creature with the body of a lion and the head of an eagle are both ideally conceivable. In order to know whether or not something is ideally conceivable, we do not necessarily have to be ideal rational conceivers ourselves. There may indeed be some cases that are marginal, and which are beyond the abilities of mere humans to settle one way or the other. Some mathematical ideas may fall into this category; but the mere fact that we are not ideal thinkers does not entail that we are never able to make judgements about ideal conceivability.

Chalmers also draws a distinction between negative and positive conceivability (2010, p144). Positive conceivability means that it is possible to imagine a scenario in full, with no gaps, so to speak, in our mental picture of what it would be like. Negative conceivability, on the other hand, is a weaker notion: it means that we can imagine the defining features, the outline of a scenario, and we can see that there is no inherent contradiction in it – but we may not be able to imagine (whether visually or otherwise) what the scenario would be like in all its detail. For an ideal rational conceiver, this distinction would not arise – a scenario would either be conceivable, or it would not. The distinction only arises for beings of finite power – beings that may not be able to imagine a scenario in full detail, but can form a rough conception of it. For a limited rational being, if a scenario is positively conceivable then it must also be negatively conceivable; but negative conceivability does not entail positive conceivability.

Now to possibility. The most fundamental type of possibility is *metaphysical* possibility. This is defined as unrestricted possibility, or possibility *simpliciter*. Metaphysical possibility and related modal notions can also be defined in terms of possible worlds: a proposition is metaphysically possible if and only if there is a possible world in which it is true; it is metaphysically necessary iff it is true in all possible worlds; it is metaphysically impossible iff there is no world in which it is true.

But what is a possible world in this context? A world is a maximally detailed way things could be, or could have been. I will take it that possible worlds exist, in the sense that we can quantify over them and they are legitimate objects of reference. However, I am not committed to the view in Lewis (1986) that they exist in the same way that the actual world – the way things actually are – exists. Rather, I take it that they exist in the actual world as mere, unactualized, possibilities. We can think of possible worlds as corresponding to world-sized properties: each possible world corresponds to the property of the world being a certain way. Only one such property is in fact instantiated by the world – namely the property of being the way things actually are. But there is an infinity of uninstantiated properties, each of which corresponds to a way things could have been, but aren't. These uninstantiated properties exist, in the actual world, as possibilities. Therefore we can also speak of the *space* of possible worlds, by which I mean the totality of uninstantiated world-sized properties – all of the ways things could have been.

We can also define various restricted types of modality. These apply relative to some subset of the (metaphysically) possible worlds. A proposition will be possible in some restricted sense in a world if and only if it is metaphysically possible and it meets the relevant restrictions that are operative in that world. For example, a proposition is physically possible in a world if and only if it is metaphysically possible and conforms to the physical laws in that world; it is causally possible in a world if it is metaphysically possible and it conforms to the causal laws in that world; and so on.

Such restrictions will not apply in the same way to all possible worlds, but will vary from one metaphysically possible world to another. If they applied equally in all possible worlds, then they would themselves be metaphysically necessary, and would not constitute a relative or restricted modality. For example, it is generally thought that the laws of physics are contingent (though this is not universally accepted, see e.g. Goff & Papineau, 2014). If this is right, then there are metaphysically possible worlds in which the laws of physics differ from those of the actual world. Thus what is physically possible in the actual world may be physically impossible in other worlds, and vice versa. In this thesis, I will not be concerned with any such restricted modality as such. The aim is solely to investigate the nature of unrestricted, or metaphysical modality.

Can we shed any further light on the nature of metaphysical possibility and necessity? One approach is to identify metaphysical modality with an idealised notion of *epistemic* modality. We can begin by defining a notion of epistemic possibility that is relative to a rational enquirer's particular subjective point of view.

Something is epistemically possible for an enquirer if and only if it cannot be ruled out on the basis of what that enquirer knows; it is epistemically necessary for an enquirer (and, *a fortiori*, possible) if its truth is entailed by what is known to them (cf Kment 2012). Thus epistemic modalities are always relative to the point of view of an enquirer, and their base of prior knowledge. If my knowledge of geography is very weak, then perhaps it is epistemically possible for me that Rome is the capital of France. When asked whether Rome is the capital of France, I will reply that, *for all I know*, it is possible. But for someone who has perfect knowledge of the capitals of Europe, it will be epistemically impossible that Rome is the capital of France. *Given what they know*, it cannot be so. Indeed, for such an enquirer, it will be epistemically necessary that Paris is the capital of France – given what they know, it must be so.

But this relative notion of epistemic modality cannot shed any light on the nature of metaphysical modality. In this relative sense, what is epistemically possible for you may not be so for me; and what is epistemically necessary from one point of view may not be epistemically necessary from another. But what if we consider the limiting case of an ideal rational conceiver who is not permitted to use any *a posteriori* knowledge as a premise in any argument? From the point of view of such a conceiver, the epistemically necessary truths are those that can be known on the basis of pure reason alone, without having to rely on any empirical premise – that is, they are the truths whose negations are not ideally conceivable.

What is the relationship between the epistemically necessary truths (for an ideal rational conceiver) and the truths that are knowable *a priori*? The answer depends on exactly how we define the *a priori*. For the purpose of this thesis, I define an *a priori* truth as one whose negation is not ideally conceivable.[5] On this definition, the epistemically necessary truths (for

[5] Therefore this notion of the *a priori* differs from Kant's. Kant's definition of *a priori* is, roughly, that something is *a priori* just if it is true for all possible experience. The definition that I will use in this thesis is narrower than Kant's: for me, something is *a priori* just if there is no ideally conceivable scenario in which it is

an ideal conceiver who is not allowed to use any empirical premises in their reasoning) are just the *a priori* ones. Similarly, the epistemically possible truths are those that are not ruled out by the knowledge of our ideal rational conceiver (who is not allowed to use any empirical premises in their reasoning). So therefore, from the point of view of an ideal conceiver who is not allowed to use any empirical premises in their reasoning, the epistemically possible worlds are just the ideally conceivable scenarios. Where the term *epistimically possible* is used in the literature, it is usually in relation to this limiting case. I only use it in this sense from now on, unless otherwise specified. Thus defined, ideal conceivability entails ideal epistemic possibility, and vice versa.

So the central question is: what is the relationship between epistemic possibility and metaphysical possibility? We have two competing views: modal monism (e.g. Chalmers 1996, 2010, Schroeter 2012, 2.3), and modal dualism (e.g. Edgington 2004). Modal monism, or modal rationalism, is the view that metaphysical modality and ideal epistemic modality are fundamentally one and the same thing. Thus every world that is epistemically possible is metaphysically possible, and vice versa. According to this view, the space of possible worlds is determined by what is ideally conceivable, and there is no distinct metaphysical modality. This is not to deny that there is such a thing as metaphysical necessity – it is just to deny that it is distinct from epistemic necessity. Modal dualism, in contrast, is the view that there are two fundamental types of modality: there is epistemic possibility, and, distinct from this, there is also metaphysical possibility. There is then a further question about the extent to which these categories overlap: there may be some epistemically possible worlds that are not metaphysically possible – and perhaps even some metaphysically possible worlds that are not epistemically possible.

A central claim of this thesis is that metaphysical possibility is identical to ideal epistemic possibility. They are one and the same thing, and modal dualism is ultimately incoherent. The defence of this claim can be broken down in to two issues. First, does epistemic possibility

false. In the *Critique of Pure Reason*, Kant argued that we can have *a priori* knowledge of the world because certain general conditions are *a priori* necessary for experience to be possible. But knowledge that these conditions obtain does depend on the empirical premise that *there is experience*, and hence it would not count as *a priori* for our present purposes. In other words, Kant's definition of the *a priori* allows for the synthetic *a priori*; mine does not.

entail metaphysical possibility? That is, does the fact that a scenario is ideally conceivable entail that there is a corresponding metaphysically possible world? Second, does metaphysical possibility entail epistemic possibility? That is, does the fact that some world, W, is metaphysically possible entail that it is ideally conceivable that W obtains? If, as I argue, epistemic and metaphysical possibilities are really the same thing, then of course the entailment will go in both directions. But the first entailment – from epistemic to metaphysical possibility – is more important for my purposes, since it allows the use of conceivability arguments to establish substantive metaphysical conclusions. For now, I will focus on the issue of whether this holds, and defer discussion of the entailment from metaphysical to epistemic possibility to the end of Chapter 4, where I will consider apparent cases of contingent *a priori* truths.

2.2 - *A Posteriori* Necessities

Truths that are necessary, but only knowable *a posteriori* (even to an ideal conceiver), appear to refute the claim that the conceivability of a scenario entails its possibility. Why is this so? Since these truths are only knoweable *a posteriori*, it follows that their negations are at least conceivable. But, since these truths are necessary – that is, true in all possible worlds – it follows that their negations are not possible. Hence the negation of an *a posteriori* necessity represents a scenario that is conceivable, but not possible. Examples of supposed *a posteriori* necessities include:

> Hesperus is Phosphorous.
>
> Water is H_2O.

It is conceivable that the Evening Star (Hesperus) and the Morning Star (Phosphorous) might not have been the same planet (i.e. Venus); and it is conceivable that the clear liquid in rivers, lakes, and human bodies could have turned out not to be H_2O. Indeed, it took great effort on the part of astronomers and chemists to discover these truths – they cannot be known from the armchair.

So surely, then, these truths are contingent? Well, no – so the argument goes – they are not contingent, because they are identities, and identities are necessary, at least when they involve referring expressions that are *rigid designators* (Kripke 1971, 1980). Rigid designators are terms that pick out in any possible world the very same object (in the case of names) or substance (in the case of natural kind terms) that they pick out in the actual world – as long as that object or substance exists in the world in question; if it does not, then the term will not refer to anything. A rigid designator may have its reference in the actual world fixed by means of a description – such as being the brightest celestial body in the pre-dawn sky, or by displaying various watery properties. But it will then pick out the same individual object or kind in any possible world (if it picks out anything at all), whether or not the object in question satisfies the description in that world – indeed, even if some other object satisfies the description instead.

Therefore, even though we can conceive of a scenario in which the brightest object in the pre-dawn sky and the brightest object in the evening sky are not in fact one and the same planet, this will not correspond to a world in which Hesperus is not Phosphorous. In the actual world, the terms ‘Hesperus’ and ‘Phosphorous’ denote the same object, the planet Venus. Moreover, they are rigid designators, because they will denote (for us) Venus in any world in which Venus exists. That, after all, is the whole point of names – they pick out objects, irrespective of whether or not the object in question happens to be the brightest celestial body in the pre-dawn sky. And there is no world in which Venus exists but is not self-identical. Therefore there is no possible world in which Hesperus is not Phosphorous. Therefore, although it is *a posteriori* that Hesperus is Phosphorous, it is a necessary truth.

Similarly, we can conceive of a world that exactly resembles our own, except for the fact that the clear, neutral liquid in streams and lakes has some chemical structure that is not H_2O – represented by the abbreviation ‘XYZ’ (as per the Twin Earth thought-experiment in Putnam, 1975). But, so the argument goes, the term ‘water’ is also a rigid designator. It picks out a type of microphysical substance, namely H_2O. It does so by means of the fact that, in the actual world, it is H_2O that instantiates the relevant watery properties. But the term will nevertheless pick out H_2O in any world in which H_2O exists, irrespective of whether or not H_2O, or some other substance instantiates the watery properties. Therefore worlds in which some other substance displays the watery properties are not in fact worlds in which water is not H_2O – rather, they are worlds in which some other substance plays the role that water plays in our world. There is no possible world in which water is not H_2O. Thus, although it is *a posteriori*, it is a necessary truth that water is H_2O.

If this analysis is right, then it would show that conceivability does not entail possibility. But is it right? I will argue that it is not. The most straightforward way to reject the idea of *a posteriori* necessities is simply to deny that there are rigid designators. In this case, the term ‘water’ will pick out whatever displays the watery properties in the world under consideration, and ‘Hesperus’ will pick out whichever object is the brightest celestial body in the evening sky in the world in question. Thus a world in which XYZ plays the role that water plays in the actual world will be a world in which water is not H_2O; and a world in which there are two bright planets – one in the morning, one in the evening – will be a world in which Hesperus is not Phosphorous.

But this does not seem like a very promising strategy, for several reasons. First, it does not seem to accurately represent how our referring expressions actually behave. It seems that some of them – especially proper names – really do behave rigidly. Of course, that is itself an empirical matter, and will depend upon how users of a language evaluate referring expressions in counterfactual situations. But this brings me to the second reason: whatever the empirical facts about how referring expressions behave (and I think they largely agree with Kripke), it will always be possible to *define* a referring expression and just *stipulate* that it should behave rigidly – there is nothing incoherent in the idea itself. But, having defined an expression that behaves rigidly, what is to stop us generating *a posteriori* necessities in the same fashion as the Kripke cases? So I will not argue that there are no rigid designators. A much better strategy would be to concede that there are indeed rigid designators, but argue that the Kripkean examples have been misinterpreted, and do not really prove what they are usually thought to prove.[6] That is the strategy that I will adopt. This strategy makes use of a theory of meaning known as *two-dimensional semantics*.

[6] It is not at all clear, in fact, that Kripke himself thought that his cases of *a posteriori* necessity proved that conceivability does not entail possibility. Indeed, in his argument against materialism, he argued from the fact that we can conceive of mental types such as pain being realised by different physical structures, to the conclusion that it is *possible* for mental types to be realised by different physical structures. However, the purpose of this section is not Kripke exegesis, but to address a widespread view that has evolved from some of Kripke's ideas.

2.3 - Two-Dimensional Semantics

The case I will make against *a posteriori* necessities is essentially that they are a trick, a sleight of hand. Statements like ‘water is H_2O’ and ‘Hesperus is Phosphorous’ are ambiguous: they have a sense in which they are *a posteriori*, and a sense in which they are necessary – but no sense in which they are both. Their apparent force comes from a failure to distinguish these two senses. In order to understand where these two senses come from, we need to understand the framework of two-dimensional semantics.

Two-dimensional semantics is a theory about the content of referring expressions, and about (at least a part of) the mechanism by which they refer.[7] One way to understand the theory is as an attempt to reconcile two opposing views of reference – referentialism and descriptivism. These are both attempts to answer the question: what is the meaning of a referring expression? The most obvious answer is that the meaning of a referring expression is just the object that it picks out – this, broadly, is the central claim of referentialism. But there are well-known reasons to think that this cannot be the whole story. These include Frege’s Puzzle (see Frege in Geach & Black (eds) 1952): if the meanings of the referring expressions ‘Hesperus’ and ‘Phosphorous’ were simply the object picked out, then it would be *a priori* that Hesperus is Phosphorous, since the two terms would mean the same thing – we could just inspect the meanings of the expressions to see that they are identical. Yet this is precisely what we cannot do. According to Frege, this shows that referring expressions have a sense – a cognitive value – in addition to their references, and that two terms with the same reference can nevertheless differ with respect to sense. Similarly, there is the problem of empty names (e.g. Bach, 1981): if the content of a name (or indeed a referring expression generally) is just the object referred to, then how can we assign meaning to names or expressions that do not refer? Yet it seems that we do. We can understand negative existential claims of the form ‘N does not exist’, and we can regard them as true. Thus the existence of the object named by ‘N’ does not seem to be a necessary condition of our ability to assign meaning to ‘N’. Therefore the meaning of ‘N’, whatever it is, cannot just be the object N.

[7] Not all versions of two-dimensional semantics claim to account for the complete mechanics of reference in all cases – mine does not. It is consistent with the basics of two-dimensional semantics that pragmatic factors may also be part of the mechanism by which reference is achieved in some cases.

In response to these considerations, we might think that the meaning of a referring expression is given by a description of some kind. This solves Frege’s Puzzle: the terms ‘Phosphorous’ and ‘Hesperus’ have the same reference, but they have different senses, given by different descriptions. It also suggests a way to avoid the problems associated with empty names and negative existential claims. We can understand terms that fail to refer precisely because we understand the content-fixing descriptions; and negative existential claims can be translated into quantificational logic. Thus the statement ‘N does not exist’ is equivalent to:

$$\sim \exists(x)\ [x \text{ is } D]$$

Where D stands for a description that constitutes the content of N. Descriptivist theories of meaning have the added advantage that they explain at least part of the mechanism by which reference is secured: terms refer to objects because the objects satisfy the descriptions associated with the terms.

But there are well known problems for descriptivism, too. First, different speakers may associate different descriptions with the same expression – how then can it have a common meaning? If you know Aristotle only as the teacher of Alexander, and I know him only as the student of Plato, then do we both have the same content for our respective ‘Aristotle’ concepts? And, if not, then how can we communicate – when we talk about Aristotle, will we just be talking at cross-purposes?

A second problem, which is especially acute in the case of proper names, is that descriptivist accounts of meaning would render certain statements analytic that in fact are not. If the meaning of ‘Aristotle’ is just ‘the teacher of Alexander’, then it will be analytic that Aristotle was the teacher of Alexander. But this is surely not analytic – it is an *a posteriori* historical fact that cannot be known just from the meanings of the terms involved.

Two-dimensional semantics (Stalnaker 1978, Chalmers 1996, 2010, Schroeter 2012) is a way of analysing the content of referring expressions that tries to reconcile these tensions. It presents a theory of meaning in which referring expressions can have two types of content: a non-rigid content (often descriptive), and a rigid content. It can be contrasted with zero- and one-dimensional semantics. In a zero-dimensional semantics (cf Schroeter 2012, 1.1.1), which corresponds roughly to referentialist theories of meaning, the content of a referring expression is simply the object which it picks out. For example, the expression ‘the President of the United States in 2019’ picks out a particular object – in this case, a man, Donald |Trump. On a zero-dimensional interpretation, the content of the expression is simply this object, Trump. Now, one might think that this is an incomplete analysis of the expression. After all, it might easily have been the case that Trump was not the President of the United States in 2019, and that another person, Hillary Clinton, was instead. In this case, the expression ‘the President of the United States in 2019’ would have picked out the object Hillary Clinton. On a zero-dimensional semantics, therefore, it would have had a completely different content to the one which in fact it has. Yet there is obviously *something* in the expression ‘the President of the United States in 2019’ that is the same in both cases. We have the same understanding of it; it is just that it happens to pick out a different object in different situations.

This leads naturally to a one-dimensional semantics (cf Schroeter 2012, 1.1.1), which corresponds roughly to descriptivist theories of reference. According to this analysis, the content of an expression is an *intension* – not to be confused with an *intention* – which is defined as a function from possible worlds to objects. That is, it provides a rule for picking out a particular object (or objects) in different possible situations. The object picked out may vary, depending on the situation, but the rule is the same. A natural way to think of such a rule is in the form of a description. This is not the only possible way; but a description gives us the dual benefit of providing a cognitive content that is stable across different possible situations, and providing a rule for picking out a particular object given the situation. So, on a one-dimensional interpretation, the content of the expression ‘the President of the United States in 2019’ is given by a rule which tells us, for any possible world, which object it refers to in that world – specifically, the object that is President of the United States in 2019. Of course, there will be many possible situations in which there is no object that satisfies the description of being President of the United States in 2019 – worlds in which the United States is a monarchy, or in which it does not exist at all. In these worlds, considered as actual, the expression would have no content at all on a zero-dimensional semantics; but on a one-dimensional interpretation, it has the same descriptive content in any world, even if it fails to refer to anything.

Two-dimensional semantics accepts the analysis of one-dimensional semantics and builds on it (cf Schroeter 2012, 1.1.1). According to two-dimensional semantics, it is possible to analyse the content of a referring expression in two ways (cf Chalmers 2010 pp146-147). The first is its primary or 1-intension. This is simply a rule or description exactly as per the one-dimensional analysis. But it is also possible to define a secondary or 2-intension for any referring term. The 2-intension consists of two separate elements: a description or rule exactly as per the 1-intension; and an indexical link to a particular possible world, which is considered as if it were actual (usually this is indeed the actual world). The 2-intension then picks out in any possible world the object which in fact satisfies the description in the indexically-linked (usually the actual) world.

It is possible to extend the notation of 1- and 2-intensions to distinguish 1-conceivability from 2-conceivability, and 1-possibility from 2-possibility. But we have to be careful how we interpret this notation. 1-conceivability means conceivability with respect to a 1-intension; 2-conceivability means conceivability with respect to a 2-intension. Similarly, 1-possibility means possibility with respect to a 1-intension, and 2-possibility means possibility with respect to a 2-intension. We must guard against thinking that 1- and 2-conceivability are two different kinds of conceivability, or – worse still – that 1-possibility and 2-possibility are two different kinds of possibility. I will return to this issue in Chapter 4.

Returning to our example: the 1-intension of the expression 'the President of the United States in 2019' functions exactly like the expression does on a one-dimensional analysis – it picks out Trump, or someone else, or no-one, depending on the situation; but the 2-intension picks out, in any possible world, the object which in fact satisfies the description of being the President of the United States in 2019 in the actual world. Thus, for any given possible world, the 1-intension picks out whatever object satisfies the relevant description in that world; the 2-intension picks out whatever satisfies it in the actual world. To bring this out more clearly, consider the sentence:

> *The President of the United States in 2019 might not have been the President of the United States in 2019.*

Is this sentence true? The answer depends on how we read it. There are two readings on which it is plainly false. If we interpret both occurrences of the expression 'the President of the United States in 2019' according to the 1-intension, then the sentence is false: it is not possible for there to be an object which is both F and not-F. Equally, if we interpret both occurrences according to the 2-intension, it is false: an object cannot fail to be self-identical. But on the most natural reading of the sentence, it turns out to be true. This occurs when we interpret the first occurrence according to the 2-intension and the second occurrence according to the 1-intension (or indeed the other way round): Trump, the object which in fact is President of the United States in 2019, might not have had this property. A world in which Clinton won the Electoral College in 2016 is a world in which the President of the United States (2-intension) is not the President of the United States (1-intension).

Similarly, a two-dimensional analysis can soften – if not completely remove – the problem of spurious analytic truths that arises when we treat proper names as descriptions. For example, let us suppose, for the sake of argument, that proper names have primary intensions that are meta-linguistic descriptions, so that the name ‘Aristotle’ has the primary content ‘the bearer of ‘Aristotle’’. In this case, will it turn out to be analytic that Aristotle is called ‘Aristotle’? According to a two-dimensional analysis, the answer is yes and no. Consider the true statement: ‘Aristotle is called ‘Aristotle’’. If we interpret the first, direct occurrence of the term ‘Aristotle’ according to its primary intension, then the statement is indeed analytic, and means something like the following: ‘If there exists a bearer of ‘Aristotle’, then it is called ‘Aristotle’’. But if we interpret the first occurrence of the referring expression in terms of its secondary intension, then we get a very different result. In this case, it will be a contingent, synthetic, and *a posteriori* truth, roughly as follows: ‘[the object ARISTOTLE] is called ‘Aristotle’’.

Another way of understanding the distinction between primary and secondary intensions is that when we evaluate a 1-intension at a possible world, we consider the world as actual – we ask, what would the description pick out, if the world were real? But when we evaluate a 2-intension at a possible world, we consider that world as counterfactual. We ask: given that the actual world is the way it is, what does the expression pick out in the alternative, counterfactual scenario?

The behaviour of primary and secondary intensions can be compared in two-dimensional matrixes (cf Stalnaker 1978 pp19-84, Schroeter 2012, 1.1.2-1.1.3). Below is a simple one showing the intensions of the expression ‘the President of the United States in 2019’ - henceforth to be known as *expression P* – for a model which contains four possible worlds, W1-W4. In this model, W1 corresponds to the actual world, and Donald Trump is President of the US in 2019. In W2, Hillary Clinton is President; in W3, Donald Duck made a late entry in 2016 and swept the Electoral College; and in W4, the American Colonies remain British and thus have no President.

P: ‘the President of the US in 2019’		World considered as **counterfactual**			
		W1	**W2**	**W3**	**W4**
World considered as **actual**	**W1**	DT	DT	DT	DT
	W2	HC	HC	HC	HC
	W3	DD	DD	DD	DD
	W4	Nil	Nil	Nil	Nil

How do we read this table? An intension is a function from possible worlds to objects. Thus it should map each possible world to the object which P picks out in that world.

The 2-intensions are most straightforward to read from the table. First we need to know which world is considered as actual – that is, which is the world to which P is indexically linked before we evaluate it at another world as counterfactual. In this case, let us take W1 as being actual – which, in fact, it is. So we look down the left hand column and find the row that corresponds to W1. This row then tells us, for each world considered as counterfactual, what object P picks out. And in each case it is the same object, namely Trump, irrespective of who happens to be President in that world. If we were to consider W2 as actual, then we would go to the next row down and see that in that case P would pick out Clinton in every possible world considered as counterfactual. Each horizontal row thus represents the 2-intension of P for a different world considered as actual, and therefore there are as many 2-intensions for P as there are possible worlds in the model.

How do we read the 1-intension from the table? Remember that the 1-intension of P requires us to take a particular description to each possible world and then determine what object in that world satisfies the description. Thus in W1 it picks out Trump, in W2, Clinton, in W3, a duck, and in W4, nothing. This is represented by the diagonal running from top left to bottom right, in which there is no distinction between which world is considered as actual and which is considered as counterfactual. So there are two important differences between 1-intensions and 2-intensions: first, 2-intensions are rigid, always picking out the same object in any world in which the object exists, whereas 1-intensions *may*[8] be non-rigid, picking out a different object depending on the circumstances; second, there are as many 2-intensions for P as there are

[8] They may also be rigid. For example, the phrase ‘the square root of four’ picks out the same number in any possible world, namely the number *2*.

worlds in the model, but there is only a single 1-intension, namely the function that, for any world W, picks out whoever is President in 2019 in W.

A further complication to two-dimensional semantics is that, as well as the distinction between 1-intensions and 2-intensions, there are different types of 1-intension. The most basic are simple descriptions: these are one-place functions that pick out whatever satisfies the relevant description in a world under consideration. Examples might be the description 'watery stuff' for the 1-intension of water, and 'teacher of Alexander' for the proper name 'Aristotle'. There are well-known problems with treating the contents of referring expressions as simple descriptions, though. In the case of proper names, we will end up with analytic statements that surely should not be analytic – such as that Aristotle was the teacher of Alexander (Kripke, 1980). A similar issue can arise when we treat the content of natural kind terms as simple descriptions. Suppose, for example, that the meaning of the concept 'mass' is specified by a description of what are thought to be the essential properties of mass – such as conforming to the Newtonian laws of motion and gravity. But then suppose that we discover that the Newtonian laws are not correct, and that the world behaves according to relativistic principles. What follows? The consequence is that nothing will satisfy the 1-intension of 'mass' – we will have to concede that mass does not exist. Thus if we define 1-intensions of natural kind terms with simple descriptions, we are liable to find that our ordinary existential claims about the world do not survive theoretical changes. Of course, these problems do not mean that 1-intensions are *never* simple descriptive functions – no doubt they often will be. But it does show that this way of understanding 1-intensions is not always adequate; we need another way of thinking about them.

A second type of 1-intension is the meta-linguistic description. This is not radically different from the first type – it is still a one-place descriptive function that picks out whatever satisfies the description in the world under consideration. But instead of listing a set of properties that are thought to be essential to the object in question – being watery, being the teacher of Alexander, and so forth – there is only one: being the bearer of the name or referring concept in question. Thus the 1-intension for water would be 'the substance called 'water''; the content of 'Aristotle' would be 'the bearer of 'Aristotle''; and so forth. In the case of proper names, this has the advantage that it does not lead to quite so many egregious false analyticities as standard versions of descriptivism – though it will still be analytic that Aristotle is called 'Aristotle', which may be problematic in itself (surely it is conceivable that Aristotle might not have been called ‘Aristotle’). In the case of natural kind terms, it is more robust in the face of theoretical change than standard descriptivism: it will pick out the stuff that we now call ‘water’, or ‘mass’, irrespective of what the science of the day has to say about the substances in question. But it also has the unwelcome consequence that the 1-intensions will pick out whatever is called ‘water’ by the natives of other possible worlds, irrespective of whether or not their ‘water’ has any watery properties.

A third option is to treat 1-intensions as two-place descriptive functions. These compare a substance or individual at a circumstance of evaluation to a substance or individual in an initial context of use for the referring term (cf Kaplan, 1989). The function is satisfied just in case the substance or individual in the circumstance of evaluation is the *same* as that in the original context of use. Thus the 1-intension of ‘water’ would be (roughly), ‘the watery substance to which I am habituated’; the 1-intension of a proper name might be given by a function of the form ‘the person I encountered on-such-and-such occasion’. The importance of this is that it is not enough for a substance or individual to have the right set of properties, or bear the right name, for it to satisfy the 1-intension. It must also be the *same* thing – type or individual – that is specified by the initial context of use. The question of exactly what makes something the same thing as that which I previously encountered is of course a difficult one. In the case of a natural kind term, it might be a matter of having the right microphysical structure. So if I encounter a watery substance that (unbeknownst to me) is not in fact H2O, it would not satisfy the 1-intension of water because it is not the *same* watery substance to which I am habituated.

But treating the 1-intension as a two-place function avoids this problem: to satisfy the intension, it is not enough to have the same properties (whatever they may be) – it must be the same individual, or the same microphysical substance. In this case, sameness might be a matter of causal-historical continuity. Thus 1-intensions of this type have an indexical link to a particular world and context of use considered as actual – and this partly determines which objects will satisfy the intention. This raises the question of how they differ from 2-intensions. Surely the content of the intension will be different, depending on which world is considered as actual? Broadly, the answer is that the content of the 1-intension is the function itself – its truth condition is just that we have the same substance or individual at the circumstance of evaluation as at the original context of use.

A fourth possibility combines elements of types 2 and 3: we can have an intension that is both comparative and metalinguistic. Hence the 1-intension of ‘water’ would be (something like) ‘the substance called ‘water’ to which I am habituated’; the 1-intension of a proper name would have the form (roughly), ‘the object at the root of the causal-historical chain of my present use of [name ‘N’]’ – that is, causal descriptivism (Loar 1976).[9]

These possible variations are summarised in the table below:

Types of 1-intension		
Type of Function	**Example – Natural Kinds (‘water’)**	**Example – Proper Names (‘Aristotle’)**
(1) Simple description	‘The watery stuff’	‘The teacher of Alexander’
(2) Meta-linguistic description	‘The stuff called ‘water’’	‘The bearer of ‘Aristotle’’
(3) Two-place descriptive function	‘The watery stuff I am habituated to’	‘The individual encountered in original context C’
(4) Two-place metalinguistic description	‘The stuff called ‘water’ that I am habituated to’	‘The object at root of causal-historical chain for my present use of ‘Aristotle’’

In some cases, it may not be clear exactly what 1-intension is associated with a referring expression. There may be some ambiguity, or it may depend upon the context. Nichols, Pinillos and Mallon (2016) present some interesting research into this phenomenon. In ancient and medieval Europe, some scholars believed in the existence of a creature called *catoblepas.* Catoblepas was said to have the body of a buffalo and the head of a boar, and to be able to kill a man with its stare. Of course, no beast matching this description has ever existed. Modern scholars believe that reports of catoblepas were based on real-life encounters with wildebeest. Now, report Nichols *et al*, when these facts are explained to people, they are often inclined to endorse both of the following statements:

a) Catoblepas do not exist.

[9] Of course, this would be a very complicated content for a proper name. But the suggestion is not that this is, in all cases, what we understand when we understand a name. The suggestion is just that this may, in some cases, be what we understand by a name – as, for example, when we understand the name ‘Aristotle’ to mean (something like): ‘The person, whoever he may have been, known to history as ‘Aristotle’.’

b) Catoblepas are wildebeest.

Of course, (a) and (b) cannot both be true (given the fact that wildebeest do exist). What is going on? One possibility (this is not quite how Nichols *et al* explain the phenomenon, but their account has similarities to my proposal) is that our concept of catoblapas is ambiguous: we have a 1-intension constituted by a simple description ('body of a buffalo, head of a boar, death-stare..'), and we also have a 1-intension that reflects our knowledge of the linguistic history of the term ('the beasts (whatever they may be) that are the historical source of the term 'catoblepas''). When confronted with statements (a) and (b), people simply select whichever 1-intension is most appropriate to the circumstances. Hence statement (a) is true if we select a type-(1) primary intension; statement (b) is true if we select a type-(4) primary intension.

2.4 - The Two-Dimensional Account of *A Posteriori* Necessity

We are now in a position to apply two-dimensional semantics to the Kripkean *a posteriori* necessities. The two-dimensional account is that the Kripke cases are ambiguous: they can be interpreted according to either the 1-intensions or the 2-intensions of the referring expressions involved. Crucially, which interpretation we give determines the modal behaviour of the resulting proposition, as well as its epistemological status. Specifically, the 1-intension interpretation is *a posteriori*, but it is also contingent; the 2-intension interpretation is necessary, but it is also in fact *a priori*. Thus the Kripkean examples do not have a reading on which they are both *a posteriori* and necessarily true.

We can illustrate this by means of an example. In the following model, W1 is the actual world, in which the colourless, pH neutral liquid found in rivers and which we call 'water' has the chemical structure H_2O. W2 is a Putnam-esque Twin-Universe, in which the colourless, pH-neutral liquid that the natives call 'water' has an alternative chemical structure, represented as XYZ. The intension-table below gives the truth-values for the statement 'water is H_2O' in the respective worlds.

‘Water is H_2O’		World considered as **Counterfactual**	
		W1	**W2**
World considered as **Actual**	**W1**	T	T
	W2	F	F

The horizontal rows represent the 2-intension interpretations of statements. If we consider our own world – in which the watery properties are instantiated by H_2O – as actual (which, of course, it is), then the 2-intension of ‘water’ is fixed as H_2O. Therefore our concept ‘water’ will pick out H_2O in all possible worlds, when considered as counterfactual. Therefore the 2-intension of ‘water is H_2O’ is true at all possible worlds – it is 2-necessary. Notice that if the actual world had been otherwise – if W2, and not W1, had turned out to be actual, and the watery properties had in fact been instantiated by XYZ – then the 2-intension of ‘water’ would have been fixed as XYZ, and it would have been 2-necessary that water is XYZ. Therefore, as the table shows, it would be false in all possible worlds that water is H_2O.

What of the 1-intension interpretation? The primary intension of ‘water’ is a function that picks out in a world whatever satisfies the function in that world. So it is (roughly) a description such as ‘the watery stuff’. So the primary intension of ‘water’ picks out H_2O in W1, and XYZ in W2. The primary interpretation of ‘water is H_2O’ is represented by the diagonal row of truth-values from top left to bottom right (cf Stalnaker 1978, p81). Hence we can see that, on the primary interpretation, it is true that water is H_2O in W1, and false in W2. So it is 1-possible that water is H_2O, and also 1-possible that water is XYZ.

In Chalmers’s language (2010, p146), if the primary interpretation of a statement S is true at a world W, then W *verifies* S; and if the secondary interpretation of S is true at W, then W *satisfies* S. Let S be the statement ‘water is H_2O’. Hence we can see from the table above that W1 both verifies and satisfies S, and W2 satisfies S but does not verify it (and, further, W2 verifies but does not satisfy ‘water is XYZ’).

Therefore, according to the two-dimensional analysis, on the primary interpretation it is both conceivable and possible that the statement ‘water is H_2O’ is false; it is only on the secondary interpretation that it is necessary that water is H_2O. In fact, we can make draw an even stronger conclusion: on the secondary interpretation, it is not even *conceivable* that ‘water is H_2O’ is false. For what would a world look like in which the 2-interpretation of ‘water is H_2O’ is false?

It would have to be a world in which H_2O is not H_2O. Therefore, on the secondary interpretation, *nothing* could count as a world in which water is not H_2O – which is precisely why it is 2-necessary that water is H_2O.

In summary: it is 1-conceivable and 1-possible that water is not H_2O; and it is 2-necessary that water is H_2O, but it is not 2-conceivable that it should be anything else. Therefore the Kripkean *a posteriori* necessities lose their apparent force. It is true that there is a sense in which it is conceivably false that water is H_2O, and there is a sense in which it is necessarily true that water is H_2O – but these senses are different. There is no unequivocal sense in which it is both conceivably false and yet necessarily true that water is H_2O.

Often a statement will be ambiguous as to which idea is being expressed. Let us consider the Hesperus/Phosphorous case. What do we mean when we say that Hesperus is Phosphorous? There are two ways of interpreting the statement. The first is that it is a statement about property instantiation: roughly, that there is in fact one object (the planet Venus) which instantiates both the properties associated with the term 'Hesperus' and those associated with 'Phosphorous'. This of course is a contingent fact about the world; it might have been otherwise, and it can only be known *a posteriori*. There are possible worlds in which there is not one object that appears both in the morning sky and the evening sky, but two distinct objects, one of which appears in the morning, and one in the evening. This interpretation corresponds to the primary interpretation, and it is therefore 1-possible that Hesperus is not Phosphorous. But on this interpretation, the statement that Hesperus is Phosphorous does not express an identity relation.

The other, secondary, interpretation of 'Hesperus is Phosphorous' is that it expresses the identity between the object which happens to bear the properties associated with Hesperus and the object which bears the properties associated with Phosphorous. On this interpretation, the statement is necessarily true – because there is no possible world in which Venus is not identical to itself. But it is also not particularly informative from an astronomical point of view.

The reason that it appears to be both necessary and *a posteriori* that Hesperus is Phosphorous is that the referring expressions involved are ambiguous. Thus when we assert that Phosphorous is Hesperus, there are two ways to interpret our assertion, which ordinary speech does not distinguish. It is conceivably false that there is only one object in both the morning sky and the evening sky; but there is no possible scenario in which Venus is not Venus. Hence we have the appearance of something which is both conceivably false and necessarily true; but in fact the thing that is conceivably false is not the same as the thing that is necessarily true.

But is two-dimensional semantics correct, though? Is it the right analysis of semantic content for referring expressions? What is certainly true is that, for at least some referring expressions, we can define both a descriptive sense and a rigid sense. Using these two senses, we can then explain what would otherwise be peculiar modal behaviour on the part of some statements. For example, there is no doubt that the phrase 'the President of the United States' can be interpreted either non-rigidly or rigidly – and, if this is not the explanation of why the statement 'the President of the United States is the President of the United States' has a reading on which it is contingent, then what is?

So the question is not whether or not two-dimensional semantics is a coherent analysis of meaning – it is. And the question is not whether or not there are some referring expressions (and therefore some statements) that conform to the two-dimensional analysis – there are. The real question is whether or not there are exceptions to the two-dimensional analysis. The issue is not whether there are some expressions for which two-dimensional semantics is true; the issue is whether there are expressions for which it is false. That is, are there expressions for which there is no associated descriptive sense? I will return to this issue in more detail in the next chapter.

Conclusion

I have argued in this chapter that the default presumption in favour of modal dualism is not justified by the Kripkean *a posteriori* necessities. If two-dimensional semantics is a viable account of the content of referring expressions, then the Kripke cases do not force us to accept modal dualism. It is possible to give a deflationary account, according to which there is a sense in which they are *a posteriori,* and there is a different sense in which they are necessary – but there is no sense in which they are both. Therefore, if the two-dimensional analysis is right, then none of these cases will offer a counter-example to the thesis that conceivability entails possibility – they will not refute modal monism.

Of course, this does not entail that modal monism is correct. There are two possible responses for an advocate of modal dualism. The first is to deny that two-dimensional semantics is a viable theory of meaning, or that it provides a viable analysis of the Kripke cases. Whilst it is beyond the scope of this thesis to explore every possible objection to two-dimensional semantics, this does not seem like a promising line of attack. Two-dimensional semantics is, at least, a viable and coherent theory of meaning. It is at least possible to interpret the Kripkean examples in accordance with it. There is nothing in the data that blocks this option.

The second response (e.g. Goff & Papineau, 2014) is to draw a distinction between two kinds of *a posteriori* necessity: there are *weak* necessities, and there are *strong* ones. The weak necessities are those, like the Kripke cases, that are vulnerable to the deflationary two-dimensional analysis. It is at least possible to read them as ambiguous, to maintain that that they do not present a single, unequivocal sense that is both necessary and *a posteriori*. But – so the argument goes – there are some *a posteriori* necessities that are not vulnerable to this deflationary account. They cannot be explained away by two-dimensional semantics; they are unambiguously both necessary and *a posteriori* – they are *strong.* They will succeed where the standard cases fail, by providing an irrefutable case that conceivability does not always entail possibility. In the next chapter, I will look at some of the main arguments for strong necessities.

Chapter 3: In Search of Strong Necessities

Introduction

In the last chapter, I argued that two-dimensional semantics gives us an alternative, deflationary way to interpret the standard examples of alleged *a posteriori* necessity. The central idea is that the standard cases – that water is H_2O, that Hesperus is Phosphorous – are in fact ambiguous: they have a sense in which they are necessary, and a sense in which they are *a posteriori* – but no sense in which they are both. If this analysis is right then the standard cases do not generate, via their negations, scenarios that are conceivable but impossible. So the standard cases are not decisive – they do not prove that conceivability does not entail possibility.

But this does not automatically mean that modal rationalism is true. The important question is whether there are any cases of *a posteriori* necessity that are immune to the deflationary two-dimensional account – that are, in the jargon, *strong*. A genuine strong necessity must present scenario, S, such that S is conceivably false but necessarily true, without ambiguity.[10] There will then be a scenario corresponding to the negation of S that is ideally conceivable but metaphysically impossible. If there are any such necessities then ideal conceivability will not entail metaphysical possibility, and some form of modal dualism must be true.

But are there any strong *a posteriori* necessities? I say there are not. In fact, I will argue that there cannot be any – that strong *a posteriori* necessities are impossible, and that modal rationalism must be true. I will set out the positive argument for my position in Chapter 4. In this chapter, though, my aim is purely defensive. Rather than presenting a positive argument in favour of modal rationalism, I will consider several cases of alleged strong necessity, and outline why they are not genuine examples. It is far beyond the scope of this thesis to consider every alleged example of strong necessity found in the literature, or to consider each case in

[10] Chalmers puts it somewhat differently – for him, a strong necessity is one that has a necessary primary intension. But this comes to the same thing, since secondary intensions, if necessary, are trivially so. The essential point is that for a strong necessity, the *a posteriori* intension and the necessary intension must be one and the same.

exhaustive detail. The aim is merely to set out some of the more interesting and important cases, and explain why I do not consider them true examples of strong *a posteriori* necessity. In the following sections, I will consider five types of case:

3.1 - Demonstratives and Indexicals
3.2 - Essential Properties of Individuals
3.3 - Metaphysical Supervenience
3.4 - Undecidable Propositions
3.5 - Necessary Objects

3.1 - Demonstratives and Indexicals

The standard Kripkean *a posteriori* necessities involve either natural kind terms – such as 'water', or concepts that clearly have some associated descriptive content, such as 'Hesperus' and 'Phosphorous'. This makes them vulnerable to the deflationary two-dimensional account, since we can easily define a primary intension for such terms in the form of a descriptive function (if it is a one-place function, then ‘watery substance’; if it is a two-place function, then ‘watery substance in my normal environment’). This enables us to generate two interpretations of the supposed *a posteriori* necessity: a primary interpretation, on which it is *a posteriori*; and a secondary interpretation, on which it is necessary.

But this suggests a strategy for generating *a posteriori* necessities that are not vulnerable to the deflationary account: this is to identify referring expressions that do not have an associated descriptive content, and to use them to generate *a posteriori* necessities. Because the referring expressions do not have descriptive contents, they are not susceptible to the two-dimensional analysis – and therefore the resulting necessities will be strong.

The most obvious candidates for referring expressions that do not have primary intensions are demonstratives and indexicals. So, for example, I could point to a sample of H_2O and say ‘*that* is H_2O’. This utterance is certainly *a posteriori* – I cannot know *a priori* that the substance in question is H_2O, as opposed to XYZ. And it also seems to be necessary, if true. There may well be possible worlds in which I (or my counterpart), in identical circumstances, point to a superficially indistinguishable substance that is in fact XYZ. But such a world would not be one in which *that* was XYZ; rather, it would be a world in which some other substance stands in place of *that*. Yet it is not obvious – and many people will resist the claim – that we can give a deflationary two-dimensional account of this situation. In order to do so, we will need an account of indexical terms like 'that' which gives them a descriptive primary intension.

So the problem is this: the two-dimensional response to the Kripkean *a posteriori* necessities is in vain, unless the two-dimensional account of meaning can be applied generally to all singular terms. Therefore we need a theory of demonstratives and indexicals that assigns them a primary intension. How might that work? It is beyond the scope of this thesis to develop and defend such an account in full, but I will outline how it could work.

The first step is to define 1-intentions of demonstratives and indexicals other than 'I' and 'now' that are descriptive functions from 'I' and 'now'. Hence *you* are the person I am now taking to; *here* is the place I am now located; *that* is the object I am now pointing at; and so on.

It would be plausible to rest here and treat the terms 'I' and 'now' as being super-rigid – meaning that their only content is determined by the particular individual or time that they designate in a given context. In this case, they would be the semantically basic building blocks of more complex indexical reference. However, I think that we should press the analysis further. While it is true that the first-person pronoun refers to you when used by you, and to me when used by me, it is also true that it has a content that transcends the identity of the individuals it refers to – a primary intension.

Hence the second step is that 'I' and 'now' have primary intensions that are functions of self-reference at the level of thoughts and utterances. Hence *now* is the time of the present thought or utterance; *I* am the thinker of the present thought. In this way we can reduce demonstratives and indexicals via descriptions to an irreducible core. It is important to caveat that I do not claim that we can eliminate demonstratives *completely*. The irreducible core, for any given subject, is their own present consciousness – *this* thought – to which the subject has direct access and can directly refer. This corresponds to the 'centre' of a possible world, from which a subject can build outwards via descriptions to other indexical facts.

With that caveat, we can use this analysis to extend the deflationary account of apparent *a posteriori* necessities to cases involving indexicals.[11] Is it an *a posteriori* necessity that *that* (said while pointing at water) is H_2O? The answer is that it is indeed an *a posteriori* necessity – but it is not a strong one. The indexical term 'that' will have a descriptive content that can be interpreted either rigidly or non-rigidly; this allows us to generate the two-dimensional analysis.

The primary intension of 'that' will pick out, in any world considered as actual, the particular individual or type that plays the relevant functional role such that I am now pointing at it. In the case of water, for the Earth-like world W1, the substance playing the relevant role is H_2O, and so the primary intension of 'that' (said when pointing at a sample of water) will pick out a sample of H_2O. In the Twin Earth-like world W2, the substance playing the relevant functional role is XYZ, and so in W2 the primary intension of 'that' (said when pointing at a sample of Twin-water) will pick out a sample of XYZ. Therefore, using the primary intension of 'that', it will be *a posteriori* – and contingent – that *that* is H_2O. The primary intension itself remains constant – it is just a function from possible worlds to objects, picking out in any world the object, if there is one, that plays the relevant functional role (that I am pointing at it). But the object it picks out will vary from world to world.

But the secondary intension of 'that' will vary, depending on which world is considered as actual. If W1 is actual, then it will designate a sample of H_2O. If W2 is actual, then it will designate XYZ. But whatever it designates, it will designate rigidly. If W1 is actual, then the secondary intension will pick out H_2O in all worlds; if W2 is actual, then it will pick out XYZ

[11] But what about the irreducible core of indexicals? What if it is true that *this thought* is identical to, say, C-fibre stimulation? In this case, we do not have a descriptive primary intension associated with the term 'this thought' – the content is just the thought itself. But then we cannot apply the deflationary two-dimensional analysis, and separate the necessary sense from the *a posteriori* sense. But, as Kripke's (1980) modal argument against materialism shows, such identities cannot be true. The reason they cannot be true is precisely because if true, they would be necessarily true – but it is possible that *this pain,* whose concept rigidly designates a particular subjective quality, could have been realised by a different physical structure. Of course, the physicalist might reply that this response begs the question of whether conceivability entails possibility, which is precisely what is at stake – but the physicalist here is also guilty of begging the question, by proposing as an example of strong necessity that physicalism is true.

in all worlds. So the secondary intension of ‘that’ functions just like Kaplan’s (1978) term of art *dthat* – it is a demonstrative term that rigidly designates a particular object.

Let us assume that W1 is actual, and that the substance I am pointing at is in fact H_2O. In that case, using the secondary intension, it will be necessary that *that* is H_2O – because here the 2-intension of ‘that’ will be H_2O. So on the two-dimensional analysis, the demonstrative ‘that’ functions like a natural kind term. We get a primary reading that is *a posteriori* and contingent, and a secondary reading that is necessary but trivial. But once again there is no interpretation of the statement ‘that is H_2O’ that is both necessary and *a posteriori*. As long as we can associate a descriptive content with indexical terms, they function no differently than natural kind terms in this respect.

This two-dimensional account of indexicals has similarities with Kaplan’s (1989) theory of indexicals – but there are differences, too. In Kaplan’s account, indexicals have both a *character* and a *content*. The character is the linguistic meaning of the term; it supplies a cognitive value that is independent of the identity of the reference. Character is the cognitive value that all uses of terms like ‘I’, ‘here’, and ‘now’ have, irrespective of the particular speaker or context of utterance. But the character of a term also plays a role in determining a specific content: it is a function from a *context of use* of the term to a content. A context of use corresponds, effectively, to a possible world considered as actual, in which the identity of the speaker or thinker is specified along with a time and place. The content, thus determined, is itself a function from *circumstances of evaluation* – effectively, a world considered as counterfactual – to the extension of the term in that circumstance. So indexical reference is a two-step process: character plus context of use determines content; and content determines reference in a circumstance of evaluation. The character of the first-person pronoun plus the context in which it is used determines a particular content – namely, the particular speaker who used it in that context; this content then picks out that individual in any further circumstance of evaluation.

At first sight, this might look two-dimensional semantics, put in different terms. The character of a term looks a lot like a primary intension, and the content looks like a secondary intension. But there is a difference. In Kaplan’s theory, character alone does not determine truth-conditions. Character plus context determines which proposition is expressed by an utterance; but character on its own expresses no proposition. A primary intension, on the other hand, does determine its own truth conditions, albeit ones that are independent of the identities of the objects involved.

So, in Kaplan's account, the character of the statement 'I am here' is not sufficient to determine any particular truth-conditions. It is only when we combine this character with information about the context of use – who the speaker is, and where they are – that we get a specific content, that person X is in location Y. In the two-dimensional account, the primary intension of 'I am here' does have a (trivially true) truth-condition: that the speaker (whoever they may be) is at their own location (wherever that may be). But, on the two-dimensional account, there is an additional, secondary content, which is identity-dependent, and which corresponds to the content in Kaplan's sense.

Why do these subtle differences matter? Because Kaplan's theory is not able to deal with apparently strong *a posteriori* necessities involving indexicals – for that, we need the two dimensional account. To show that statements like 'that is H_2O', or 'I am a human being' do not constitute strong necessities, we need to be able to show that they are ambiguous – that they express a proposition that is *a posteriori* and contingent, as well as a proposition that is necessary but trivial. But this means that they must express two propositions simultaneously – they must have two sense, both of which determine truth conditions. Kaplan's notion of character is not up to this task; but a primary intension is.

We can see how the two-dimensional account deals with apparently strong *a posteriori* necessities involving the first person pronoun. Consider the following scenario. I find myself thinking, and wonder about the identity of the thing that is doing the thinking. I think: 'What sort of thing am I?' Now, it appears to me that I am a particular human animal[12], with a certain

[12] Or Lockean person, depending on one's view on the primary bearer of personal identity. I am in fact more sympathetic to Lockeanism than animalism, but I will not take a side in that debate here. Equally, there may well be possible situation is which it is not at all clear what the reference of 'I' should be, because it is not clear what object the thinker of the present thought really is, or where it is located (see e.g. Dennett 1978, 'Where am I?'). In general, there are many further questions about what, in the actual world, is in fact the thinker of my present thoughts, and about how this object (whatever it may be) does or does not persist through time, and whether it would survive in various counterfactual possibilities. But all these questions about personal identity are tangential to the present point, which is the distinction between primary and secondary intensions of the first-person pronoun.

history, and so forth. But perhaps this appearance is illusory – perhaps I am in fact a machine, a cyborg or similar, and all these memories and impressions of being human have been implanted in me to deceive me. So we have two possible worlds: in world W1, the thinker of the occurant thought 'What sort of thing am I?' is a human animal, with (mostly) veridical memories and impressions of being such; in world W2, the thinker of this thought is in fact a cyborg, with illusory impressions of being human.

So what is the content of the pronoun 'I', as used in this instance? The primary intension of 'I' is the descriptive function 'the thinker of this thought', and, in any world considered as actual, it will pick out the object which happens to be the thinker of the thought in that world. Thus in world W1, the primary intension of 'I' picks out a human being; in world W2, it picks out a cyborg, and so forth. But the secondary intension of 'I' will rigidly designate the object that is the thinker of *this thought* in the actual world. So, let us suppose that we are in W1, that I am in fact a human animal, and that this animal is the thinker of my present thoughts. In that case, the secondary intention of my use of the first-person pronoun will rigidly designate that particular animal. In the secondary-intension sense, it will not be true that in W2 I am a cyborg, because an animal cannot be a cyborg. What will be true on the secondary intension reading is that in W2, there is a cyborg thinking subjectively identical thoughts to me in W1. Similarly, there will be worlds in which the human animal that (on this hypothesis) I am identical to exists, but does not think *this thought* or anything resembling it. But that is not a problem: it is till true of me in such worlds that I am (identical to the object that is) the thinker, in the actual world, of *this thought.*

On this analysis, the proposition 'I am a human being' turns out to be an *a posteriori* necessity – but not a strong one. It is susceptible to the same two-dimensional analysis as cases involving natural kind terms, and cases involving demonstratives. So there is certainly a sense in which it is *a posteriori* that I am a human being – it is conceivable that I am not; and perhaps, one day, I will be shocked to discover that I am in fact a cyborg. And there is also clearly a sense in which it is necessary that I am human: whilst there are possible worlds containing cyborgs that are psychologically and subjectively indiscriminable from me, given that I am in fact a human, none of them can be *me*. But – and this is the crucial point – the sense of 'I' in which it is *a posteriori* that I am a human is different from the sense of 'I' in which it is necessary that I am human. Therefore it is *a posteriori* and contingent that [primary intension]-I am human. But, if W1 is actual, then the secondary intension of ‘I’ will be fixed such that it is necessary (and indeed *a priori*) that [secondary intension]-I am human.

The two-dimensional analysis of indexical terms tends to support a deflationary view of indexical facts. The idea is that all demonstrative and indexical terms can be analysed as descriptive functions of the first-person pronoun and the present time; and the indexicality of the first-person pronoun and the present time is a function of self-reference at the level of individual thoughts.

How does this deflationary account work in the case of individuals? Are there metaphysically basic facts about identities? Intuitively, it might seem that there are. It seems to me a substantive fact that I am *me.* Could I not have been someone else? As I have outlined above, there is a sense in which I could have been someone other than me: my present consciousness (or its counterpart, if any is identifiable) could have been had by a different object than the particular human animal that is its actual bearer; equally, my human animal or Lockean person might have been the bearer of a very different consciousness in the present time, had history taken a different path. But this sense is adequately captured by the two-dimensional analysis of the first-person pronoun; it requires nothing more than the two-dimensional framework and self-reference at the level of thoughts.[13]

[13] But is there a possible world that is a minimal physical and mental duplicate of the actual world, but which differs with respect to the facts about individual identities? Is there a world that is exactly the same as the actual world, except that I am occupying the place of your consciousness, and you occupy mine? We are not here considering the possibility that your human animal is thinking the thoughts that in the actual world are thought by my animal, and vice versa; here we are interested in the much more radical hypothesis that all of the psychological and animal facts might be exactly as they are in the actual world, but the facts about identities are changed.

We can imagine that thinking subjects have souls, each of which has a uniquely identifying mark – a serial number, visible only to God – and that these can be transposed between subjects without changing any of the facts about consciousness. In fact, it is conceivable that the souls are switching between the various streams of consciousness in that actual world all the time, and we would never know – it would make no difference to us. Not being a verificationist, I do not think this is a meaningless hypothesis; it is metaphysically possible. But it is nothing to do with indexicality as such. Rather, it is about the identity conditions of the reference of the first-person pronoun – and as a hypothesis it constitutes an alternative to animalism or Lockeanism, rather than a theory of indexicality. What is important about indexicality is how the world looks and feels from the point of view of particular subjects;

In summary, indexical facts do exist, and there are substantive truths associated with them, including weak *a posteriori* necessities. But the two-dimensional analysis shows that they do not in fact yield strong necessities. Moreover, indexical facts supervene metaphysically on the totality of physical and psychological facts. They are reducible via description to facts about the identities of speakers or thinkers and which time is the present. We can give then a deflationary account of them, and of any associated *a posteriori* necessities, using the two-dimensional analysis of the basic indexical terms *I* and *now*.

Given that indexical facts are not metaphysically basic, we can think of them in terms of *centred worlds* (Chalmers 2010), where a centred world is:

> […] an ordered triple of a possible world and an individual and a time in that world. [p546]

My analysis goes a step further than Chalmers's: like him, I reduce all other indexicals and demonstratives to individuals and times; but I further claim that individuals and times themselves are reducible via description to the identity of a subject's present consciousness. The irreducible demonstrative core is the identity of *this thought* – to which, being immediately present to in consciousness, we refer directly. The identity of *this thought* centres the world.

The centring of a world does not require the addition of any metaphysically basic facts about individuals, or times. Rather, it is like putting a pin in a map of the universe – a marker, indicating that *I am here (now).* The rest of the world can then be described in terms of how it is related – temporally, spatially and so forth – to this fixed reference point. So a centred world is not a world plus an indexical fact; it is a world as described from a particular point of view within the world. And, since the point of view is just that of an individual at a time, it is not something over and above the totality of physical and mental facts that the world contains.

This discussion of the metaphysical status of indexical facts has strayed some way from the original point of this section. But the fundamental point is that there is no good reason why the two-dimensional analysis cannot be extended to demonstratives and other indexical terms. Therefore apparent *a posteriori* necessities involving indexical terms are just as vulnerable to

what matters is the presentness of now, and the self-awareness of the first-person perspective. But these phenomena supervene metaphysically on the physical and psychological facts of the world.

the deflationary two-dimensional analysis as the more standard ones involving natural kind terms. So there is no reason to think that indexical terms have any special power to generate strong *a posteriori* necessities.

3.2 - Essential Properties of Individuals

Goff & Papineau (2014) acknowledge that the Kripkean weak necessities do not disprove modal monism, and propose several examples of supposed strong *a posteriori* necessities. The first type of strong necessity supposedly arises from the possibility of what they term 'radically opaque terms'.

Goff & Papineau argue that the two-dimensional analysis of *a posteriori* necessities is only possible because:

> […] for any *a posteriori* necessity, the terms involved have extra content in addition to their referent. 'Water' does not just refer to H_2O, but expresses the property of being the colourless, odourless liquid in oceans and lakes. 'Hesperus' does not just refer to Venus, but expresses the property of being the heavenly body visible in the evening. It is these extra contents that allow us to construct 'surrogate' possibilities [...] alongside the necessities […] [Goff & Papineau 2014, 1.2; Authors' emphasis]

This is quite right. The 'extra content' – as Goff & Papineau acknowledge – is just the primary intension of the referring expression, and it is this that gives rise to the 1-possibility that, for example, water is not H_2O. But at this point their analysis takes a wrong turn. They ask why there should not be terms that refer directly to their referents, without any extra content – that is, why there should not be terms which have no primary intensions, only secondary intensions. They call such expressions 'radically opaque terms'. They then argue that 'If there are two distinct but co-referring radically opaque terms, then putting them together with an identity sign between them would give rise to a strong necessity'. They speculate that 'Cicero' and 'Tully' might be radically opaque terms, and that therefore 'Cicero is Tully' might express a strong necessity.

Using this idea, Goff & Papineau argue (at 1.2) as follows (I am paraphrasing): Let us suppose, for the sake of argument, that 'Cicero' and 'Tully' are both radically opaque terms, and that in fact Cicero is Tully. Then the falsity of 'Cicero is Tully' would be conceivable, because there is not enough information in the content of the respective terms to determine *a priori* that Cicero is Tully; and therefore we can conceive of the possibility that they are not identical. And yet, since there are no primary intensions, there is no possibility that Cicero is not Tully.

There are two problems with this argument. The first is that it is not credible in my view that proper names of external objects are radically opaque. The only possible radically opaque terms are phenomenal concepts. This is because any radically opaque terms that had external objects as their contents would lead to strong mental externalism, a position which I reject (see Chapters 6 and 7).[14] But I will put this objection to one side for now. The more significant problem is that even if 'Cicero' and 'Tully' were radically opaque, this would still not generate a strong necessity. Suppose, for the sake of argument, that 'Cicero' and 'Tully' are indeed radically opaque terms, and their only content is the object that they refer to, namely the man Marcus Tullius. Clearly, it will be necessary that Cicero is Tully. It is not possible that Marcus Tullius is not identical to Marcus Tullius. But nor would it be conceivable that Cicero is not Tully. Of course, if we do not know that the terms 'Cicero' and 'Tully' both refer to Marcus Tullius, then we cannot rule out *a priori* that 'Cicero is not Tully' expresses a truth. But this is merely because, if the referring expressions are radically opaque, then we will not know their contents – we will not know what it means to assert that Cicero is or is not Tully. Conceiving that 'Cicero is Tully' expresses a truth is not the same as conceiving that Cicero is Tully. To do that, we would need to know the contents of both referring expressions. But if these terms are radically opaque, and both refer to Marcus Tullius, then the contents entail *a priori* that Cicero is Tully.

In fact, I do think there is a sense in which it is conceivable that Cicero is not Tully – but that is precisely because I believe referring expressions always have identity-independent cognitive content. The fundamental problem for Goff & Papineau's proposal is that this 'extra' content of referring expressions is needed not only to generate the surrogate possibility, which they reject, but the counterfactual conceivable scenario, which they need. But they can't have one without the other. If there really are radically opaque terms that lack primary intensions, then

[14] In theory, I think it is possible that some non-phenomenal terms have *linguistic* contents are radically opaque, since I am not opposed to externalism with respect to linguistic content. But that is not relevant in the present situation, because the issue is whether it is *conceivable* that Cicero is not Tully – i.e. whether the identity follows *a priori* from the associated mental contents.

the absence of a primary intension would indeed deprive us of the counterfactual possibility – but it would also deprive us of the ability to conceive of a counterfactual scenario.

There are three lessons here. The first is that the 1-possibility of a scenario and its 1-conceivability go hand in hand. If there were radically opaque terms, then we could not hope to jettison the 1-possibility whilst keeping the 1-conceivability. All that would remain would be the secondary intensions, and the corresponding notions of 2-conceivability and 2-possibility. The second moral is that a scenario might *appear* conceivable because we are not in a position to know the contents of the terms involved – and yet, when we do know the contents, it may become apparent that the scenario is not *really* conceivable. This is liable to be the case where secondary intensions are involved. The third lesson is that this shows us exactly what is wrong with radically opaque terms. They are just Millian names – all reference, and no sense. But it is precisely because we *do* attach a sense to them before we learn their reference that it can be an interesting discovery that Hesperus is in fact Phosphorous, or Cicero was Tully. So I think the notion that there are radically opaque terms is implausible. But, even if there were, they would not give rise to strong *a posteriori* necessities.

In the same paper, Papineau (without Goff) goes further than merely claiming that *some* metaphysical necessities might not fit the two-dimensional scheme, arguing that (in his words, section 3) 'Modality is Nothing to do with Conceivability'. There are, he argues, examples of metaphysical necessity that are simply not grounded in conceivability at all. For example, it is necessary that a person should have the parents which they do in fact have, but it is not necessary that a person should have been born in the place that they were in fact born in. And this difference does not, he argues, seem to be grounded in facts about conceivability at all. So the statement 'David Papineau's father is Owen Papineau' is necessary; 'David Papineau was born in Como' is not. Yet in both cases the counterfactual is perfectly conceivable. It is conceivable that David's father might not have been Owen; and it is conceivable that he might not have been born in Como. But in one case, the conceivable scenario corresponds to a possibility; in the other it does not. As Papineau puts it:

> In the absence of some further story here, it looks as if the explanation of metaphysical modality in terms of secondary intensions, far from being grounded in facts of conceivability' is simply tracking some independently given structure of metaphysically modal facts. [Papineau in Goff & Papineau 2014, 3.2]

What should we make of this? Suppose we concede that there is some sense in which it is necessary that David's father is Owen, and that it is not necessary that David was born in Como. We seem to have two possible ways of explaining this: either there is some primitive metaphysical modality at work, such that it is metaphysically necessary that a person has the parents that they actually have; or there is some peculiarity in the way our concepts of persons function, such that there is an asymmetry between conceiving of a person having different parents than their actual parents, and conceiving of them having a different birthplace than their actual birthplace – and that the first is in some sense *inconceivable.* If modal monism is true, then the second possibility must be correct. But this presents a difficulty. As Papineau puts it:

> But how then does metaphysical modality get tied to conceivability? What does conceivability have to do with the impossibility of my having a different father, but not a different birthplace? Even though both are initially conceivable, is there some further sense in which my having a different father than Owen is nevertheless less conceivable than my having a different birthplace than Como? [Papineau in Goff & Papineau 2014, 3.2; Author's emphasis.]

As a statement of the problem, this is exactly right. If modal monism is correct, then the impossibility of David's having a father other than Owen must arise from the *inconceivability* of that scenario; and there must be an asymmetry with the alternate birthplace scenario. So how is *that* supposed to work, given that it just as conceivable that David was sired by someone other than Owen as that he was not born in Como?

One possibility is that, if we do not know either David's birthplace or his father, then it is conceivable that they are other than Owen or Como, respectively – but that once we know the relevant facts, then 'different fathers are rendered inconceivable in a way that different birthplaces are not' (as Papineau puts it). Papineau considers this idea, and gives it short shrift, pointing out that the supposition that his father is Owen does indeed rule out the possibility that it should be someone else – but, in exactly the same way, the fact that he was born in Como rules out his being born anywhere else. There is no asymmetry here.

But this dismissal is too quick, in my view. The point is not just that knowledge of the *a posteriori* facts will rule out alternate scenarios. Of course there is a sense in which it is inconceivable that David's father is not Owen, given that David's father is Owen – but that does not get us very far. While it is obviously inconceivable that {David's father is not Owen, given that David's father is Owen}, this does not mean that, given that David's father is Owen, it is inconceivable that {David's father is not Owen}. The same could of course be said of David's birth place. But I think this is not the whole story. There is something going on conceptually in the case of David's parentage that is not present in the case of his birthplace.

I think the point is rather that there are some *a posteriori* facts which serve to fix the secondary contents of some of our concepts, and that our concepts of individual persons and the *a posteriori* facts about their parentage fall into this category, but the *a posteriori* facts about their birthplaces do not. Before I know the facts about David – either where he was born or who his father was – I still know that I will consider the identity of his father to be an essential property of David, but not the location of his birth. In this respect, David's parentage is exactly like the microphysical structure of water. It is not part of our primary content, but it serves to fix the secondary content of our concept.

This, in essence, is the response that Chalmers makes (2014, pp3-9). This response is consistent with the strategic goal of modal monism to analyse metaphysical necessities in terms of a single space of epistemically possible worlds, combined with the way that certain of our concepts function in relation to certain *a posteriori* facts, plus certain *a posteriori* facts themselves. In other words, our concepts of individuals have both primary and secondary intensions, just as natural kind concepts do. The primary intension of my concept 'DP' is something like a mode of presentation of the object DP – *philosopher, called 'David Papineau'*, and so forth. The secondary intension of my concept will track some – but not all – of the actual facts that lie behind and explain this mode of presentation. So it will track parentage, but not birthplace.[15] So, if it turns out that David was actually fathered by Owen, then this fact will contribute to fixing the 2-intension, even though it is not part of the 1-intension. Just as it is 1-conceivable but not 2-possible that water is not H_2O, so it will be 1-conceivable but not 2-possible that

[15] Why one but not the other? One possible explanation, favoured by David Papineau (see Godman, Mallozzi & Papineau, forthcoming) is that many of our concepts track what he terms *super-explanatory properties*. A super-explanatory property is a property that explains a large number of the observable properties of an object or type. For example, having the chemical composition H_2O is super-explanatory of water's observable properties, and a human individual's parentage is super-explanatory of their physical properties, whereas a person's birthplace is not. I am very sympathetic to this view, but would express it in two-dimensional terms: for these concepts, we have a primary intension that tracks certain observable properties, and a secondary intension that tracks whatever turns out to be super-explanatory of those observable properties.

David's father is not Owen. In other words, there is nothing special about the present case: parentage is a case of Kripkean weak *a posteriori* necessity, and nothing more.

In this manner, we can conceive of a world in which an exact doppelganger of David Papineau, fathered by Owen, is born in some place other than Como. We have no difficulty in declaring such a world possible; its possibility follows straightforwardly from its conceivability. So it is ideally epistemically possible that David might have been born somewhere other than Como, and there is a possible world which corresponds to this scenario. We can also conceive of a scenario in which an exact doppelganger of David Papineau – born in Como – was fathered by someone other than Owen. This scenario presumably corresponds to a possible world, and so it is both conceivable and metaphysically possible. But would the doppelganger be *David*? This, it seems to me, is the crucial difference. Once we know that, in fact, David's father is Owen, we may not be willing to count any counterfactual individual as *being* David – however closely it resembles him – unless that individual is fathered by Owen.[16] Our knowledge of the actual David's parentage does not constrain what can go on in counterfactual worlds – it does not prevent there being one in which a doppelganger of David Papineau comes into being by miraculous materialisation – it merely constrains how it is proper to describe those worlds. In contrast our knowledge of the actual David's birthplace does not place any such constraint on us; it does not serve to fix any part of the content of what we mean by our concept of the person. So, in the parentage case, it is 1-conceivable and 1-possible that David's father is not Owen; but it is neither 2-conceivable nor 2-possible. In contrast, in the birthplace case, it is both 1-conceivable and 2-conceivable, and therefore 1-possible and 2-possible, that David was not born in Como. So it is possible to account for the modalities of parentage and trans-world personal identity in terms of a single space of possible worlds.

We can make a further argument for this two-dimensional account, along the following lines. If we accept that individuals can have essential properties then, whenever we are ignorant of

[16] On the other hand, some people might be willing to allow that such an individual really would be David. This seems to me to indicate that what is at stake is just a matter of how concepts of personal identity function – that is, whether or not we regard an individual's parents as being essential to their identity – and that different people may quite legitimately have different views on this, since it is an arbitrary matter. If, on the other hand, Papineau is right and there is some pre-existing metaphysical fact about personal identity, then people who think that such an individual would count as David have simply got their metaphysical facts wrong.

the essential properties of an individual in front of us, we will be confronted with a masked man scenario: we will not know *which* individual it is. For example: there is a man in front of me; I do not know who his father is. There is one possible world in which a man exactly like this one has father A; there is another possible world in which a man exactly like this one has father B. If we assume that the identity of a man's parents are an essential property of the individual, then it follows that it is not the same man in the A-world as in the B-world. But I do not know which world I am in – and therefore I do not know *which* individual this is.

Note that this problem arises whichever view we take of essential properties – whether we regard them as being grounded in metaphysical necessities, or take the two-dimensional view that they are grounded in the way certain concepts function. In either case, we must accept that there is a sense in which I do not know which individual – the A-individual or the B-individual – is the one in front of me. But this situation is not problematic for the two-dimensional view. On that view, it is just that I do not know 2-intension of my concept of the individual – just as, if I am ignorant of chemistry, I do not know the 2-intension of my concept of water. I do not know which man this is just in the sense that, if asked to consider the B-world, I would not know whether it is proper to count the B-individual as being the same man as this man. But there is still a sense in which I do know the identity of the man in front of me: I know that the individual in front of me satisfies the 1-intension of my concept. I know that it is a man called so-and-so, with such-and-such properties, and so forth. However, the situation is more problematic for the rival view. That view does not have the option of saying that there is a sense in which I know which individual is in front of me, and a sense in which I do not. It must concede that I do not know which individual this is, *simpliciter*. Maybe that is a price that an advocate of modal dualism would accept. But the important thing is that it is not possible to have it both ways. If essential properties of individuals are a matter of strong *a posteriori* necessity, then an inevitable consequence is that if I do not know the essential properties of an object, then there is no sense in which I know which object it is.

3.3 - Metaphysical Supervenience

Does metaphysical supervenience require *a priori* entailment? Block & Stalnaker (1999) argue that it does not. In particular, they argue that the totality of microphysical facts about the actual world does not entail *a priori* the totality of ordinary macroscopic physical facts about the world. But, they argue, the microphysical facts are still metaphysically sufficient for the ordinary macroscopic physical facts. It would be absurd to think otherwise – we are all physicalists about ordinary physical objects, after all. So there is no metaphysically possible world that is a minimal physical duplicate of the actual world, but which differs with respect to the ordinary macroscopic physical facts. But, because there is no *a priori* entailment from the former to the latter, such a world is ideally conceivable.

If Block & Stalnaker are right, then it is a strong *a posteriori* necessity that the ordinary macroscopic physical facts about the world supervene on the microphysical facts. This would constitute a genuine counter-example to modal rationalism, and would show decisively that ideal conceivability does not entail metaphysical possibility. Moreover, it would also mean that consciousness is not a special case when compared to other macroscopic phenomena. Chalmers & Jackson, in Chalmers (2010), disagree. They argue that there is in fact an *a priori* entailment from the microphysical facts to the ordinary macroscopic ones. Who is right?

The central idea of Block & Stalnaker's argument is that *a priori* entailment requires a conceptual analysis of the higher-order facts in terms of the lower-order facts. That is, it would require ordinary macroscopic descriptions of the world to be analysible in terms of the underlying microphysics. But, so they argue, no such analysis is possible. And therefore there is no *a priori* entailment from microphysics to macroscopic facts. But, they argue, there is a metaphysical entailment; and so metaphysical supervenience does not require *a priori* entailment. Thus we can summarise Block & Stalnaker's argument as follows:

1) In general, *a priori* entailment from the A-facts to the B-facts requires a conceptual reduction of the B-facts to the A-facts.

2) No conceptual reduction of macroscopic physical facts to the microphysical facts is possible.

Therefore:

3) There is no *a priori* entailment from the microphysical facts to the macroscopic physical facts.

But:

4) The macroscopic physical facts supervene metaphysically on the microphysical facts.

Therefore:

5) Metaphysical supervenience does not require *a priori* entailment.

Consider, for example, the ordinary macroscopic facts about the distribution of water in the world. The relevant A-facts are those regarding the distribution of the microphysical substance H_2O. Now, as per premise (2), there is no conceptual analysis of facts about water in terms of facts about H_2O. After all, it is not *a priori* that the colourless, odourless substance in the rivers and oceans has that particular microphysical structure; and moreover it is metaphysically possible that the same watery macro-properties could have been instantiated by some other microphysical substance. But if (1) is correct, then the corresponding version of (3) follows: there is no *a priori* entailment from the microphysical facts about the distribution of H_2O to the macro-facts about the distribution of watery properties. (Of course, if the term 'water' is interpreted according to its 2-intension, then there is a trivial entailment from facts about H_2O to facts about water, so we must separate meanings carefully). However, the corresponding version of (4) is surely true: the macro-facts about the distribution of watery properties surely supervene on the microphysical facts about H_2O. We are all physicalists about water, after all. But then the conclusion, (5), follows: we have metaphysical supervenience without *a priori* entailment – we have a strong necessity.

I take it that the argument is valid. Chalmers & Jackson reject the conclusion, (5). However, they do accept a version of (4) – that is, they accept that the macroscopic facts supervene on the microphysical facts (plus some other elements to complete the supervenience base). So Chalmers & Jackson resist the conclusion by rejecting (3). Their view is that the macroscopic facts do indeed supervene *a priori* on the microphysical facts.[17]

[17] To be more precise, Chalmers & Jackson argue that we need to include several other terms in the supervenience base of the macroscopic facts, M. The first addition (as per Chapter 1, Section 1.4, above) is a 'that's all' clause, represented by *T*, in order to rule out blockers and epiphenomenal ectoplasm. The second addition, represented by *I*, stands for the conjunction of all the indexical truths – for example that I am individual,

In order to reject (3), Chalmers & Jackson must reject either (1) or (2). But which? They accept premise (2), and I think they are right to do so. It strikes me as a truism that ordinary macroscopic concepts cannot be conceptually analysed in terms of microphysics. Even if some form of analytic functionalism is true of macroscopic phenomena, this means only that macroscopic concepts can be functionally analysed at the macro-level. It is a further, empirical claim that the microphysical facts of the universe play the relevant functional roles and thereby instantiate the macro-properties. And this seems to leave open the possibility that the macro-

i, that it is now time *t*, and so forth. However, for the sake of simplicity, I will not show T in the supervenience base, although it should be understood that it is there implicitly; and, I will exclude the indexical term *I* from the supervenience base, in accordance with my view that indexical facts are not metaphysically basic.

Chalmers & Jackson also include Q in the metaphysical supervenience base of M, where Q represents the totality of facts about consciousness. Thus (excluding T and I) we have the following entailment from the metaphysical supervenience base to the macroscopic facts:

P + Q ➔M

But surely it is question-begging to include Q in the supervenience base of M (cf Levine 2011)? If Q is part of the supervenience base of M, then P on its own does not metaphysically entail M. But that seems to undercut the whole argument between Block & Stalnaker and Chalmers & Jackson. Both sides are supposedly agreed on the metaphysical entailment from the microphysical facts to the macrophysical ones – the issue is whether this requires an *a priori* entailment.

However, whilst it might seem question begging, this is not really the case. This is because M here represents the totality of the macroscopic facts – but this will include facts about conscious objects such as human beings, and also plausibly macroscopic properties that non-conscious objects have in virtue of their relation to conscious beings. But if physicalism is false, then these elements of M will not supervene metaphysically on P alone.

But the present dispute concerns the nature of the entailment from the microphysical facts (plus T and I) to the macroscopic facts *minus consciousness.* We can write this as follows:

P ➔ M - Q

In order to simplify matters, I will hereafter take M to designate the macroscopic facts *minus consciousness*. Hence the macroscopic facts in totality, inclusive of consciousness, would be written as [M + Q]. The exam question concerns the nature of the entailment from P to M, where M represents the macroscopic facts minus consciousness. Block & Stalnaker claim that this entailment is metaphysically necessary, but not *a priori*, and is therefore a strong necessity; Chalmers & Jackson claim that it is *a priori*.

phenomena could have been realised by different micro-facts, which seems exactly right. But a conceptual analysis of the sort rejected by both Block & Stalnaker and Chalmers & Jackson would rule this out. If there were a conceptual analysis of macro phenomena in microphysical terms, then there would be an *a priori* entailment from macro to micro – which neither they nor I think is remotely plausible.

So both Block & Stalnaker and Chalmers & Jackson are surely right that, in general, there is no *a priori* reduction of ordinary macroscopic (non-conscious) facts to microphysical facts. That is, there is no conceptual analysis of macroscopic concepts in terms of microphysical concepts. There is no *a priori* entailment from facts about water (when understood in terms of its 1-intension) to facts about H_2O. Moreover, macroscopic facts are very often multiply realisable by different microphysical facts. The watery facts could have been realised by H_2O or XYZ. So premise (2) is not in question.

Thus Chalmers & Jackson's strategy for rejecting (3), and thereby resisting the conclusion (5), is to reject premise (1). So the issue turns on whether or not, in general, *a priori* entailment from the A-facts to the B-facts requires a conceptual analysis of B-concepts in terms of A-concepts. This is the heart of the argument between Block & Stalnaker on one side and Chalmers and Jackson on the other. Thus Block & Stalnaker write:

> The point of view that we are criticizing depends on […] claims that we are accepting: […] that there is no contradiction or incoherence in the extreme zombie hypothesis, the idea of a microphysical duplicate of one of us but with no consciousness […]. These ideas are supposed to show that the facts of consciousness are not *a priori* entailed by the microphysical facts. But we have been arguing that the facts about *water are not a priori entailed by the microphysical facts either*. To derive 'The earth is 60 percent covered by water' from microphysics, we need the a posteriori (necessary) truth that water = H_2O. [p 29; authors' italics.]

But Chalmers & Jackson argue to the contrary:

> Overall, [the microphysical supervenience base] implies complete information about the […] structure and dynamics of macroscopic systems and objects in the world, their spatiotemporal distribution and microstructural composition […]
>
> 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. [p 224]

In general, does *a priori* entailment from the A-facts to the B-facts require a conceptual reduction of the B-concepts to the A-concepts? And, specifically, can there be an *a priori* entailment from the P-facts to the M-facts, without a conceptual analysis of M-facts in P-terms? We can imagine a God-like intellect, given complete knowledge of the microphysical state of the universe. The question is then whether such a God-like intellect would be able to deduce the (non-conscious) macroscopic facts from this complete microphysical description. Would such a conceiver be able to know all the facts about – for example – hydrology, living beings, weather systems, astronomy, and everyday physical objects? Chalmers & Jackson say yes; Block & Stalnaker say no.

Chalmers (2012) argues for what he terms the *scrutability* of the world. An intuitive version of the scrutability thesis states that the metaphysically basic truths of the world *a priori* entail all of the truths. If A represents all the metaphysically basic truths of the actual world, and B represents all the truths of the actual world, then an ideal rational conceiver can know on the basis of pure reason alone that *if A, then B*. As Chalmers puts it:

> *A Priori Scrutability*. There is a compact class of truths such that for all true proposition *p*, a Laplacean intellect would be in a position to know a priori that if the truths in that class obtain, then *p*. [2012, p xvi; author's italics]

(The compact class of truths in question is the metaphysically basic truths; a Laplacean demon is an intellect sufficiently powerful that the world is scrutable as specified.)

The scrutability thesis is closely related to modal rationalism. If modal rationalism is true, then the world must be scrutable. How so? We start with the premise that for any truth *p*, either *p* is metaphysically entailed by the metaphysically basic truths, or *p* is itself a metaphysically basic truth. This is just a definition of what it means to be metaphysically basic. But if modal rationalism is true, then metaphysical entailment is *a priori* entailment. So, for all truths *p*, either *p* is metaphysically basic, or *p* is *a priori* entailed by the metaphysically basic truths – in which case, the world is scrutable. (Moreover, as I shall argue in the next chapter, modal rationalism is itself an *a priori* necessary truth. This entails that it is not just the actual world that is scrutable, but all possible worlds.)

Does the reverse hold? If the world is scrutable, does this entail modal rationalism? Not quite. Now if, as Block & Stalnaker claim, there is *a posteriori* metaphysical supervenience, then the world will not be scrutable. If they are right, there will be truths – the metaphysically supervenient higher-order facts – that cannot be known, even by an ideal rational conceiver, just on the basis of the lower-order metaphysically basic facts. Therefore the scrutability thesis does entail that there is no *a posteriori* metaphysical supervenience. Scrutability entails that there are no metaphysically non-basic truths as a result of strong *a posteriori* necessities. So, the scrutability thesis rules out strong necessities of the form: if [metaphysically basic fact], then [metaphysically non-basic fact]. But, at face value at least, it leaves open the possibility that there are strong necessities that unconditionally assert metaphysically basic facts. And therefore the scrutability of the world does not rule out a strong necessity asserting that God, or energy, or some other supposed metaphysically basic reality exists. So the scrutability of the world does not entail modal rationalism, although it does place significant restrictions on what strong necessities there could be.

If the world is indeed scrutable, then the really interesting questions will concern what the metaphysically basic truths are, and how the remaining truths are *a priori* entailed by them. This is the project of Chalmers (2012). One of the central issues is whether scrutability requires *definability*, where:

> *Definability*. There is a compact class of primitive expressions such that all expressions are definable in terms of that class. [2012, p3; author's italics]

Does scrutability require definability? It we want the world to be scrutable then it had better not, because, as we have seen, definability is not generally true. For example, water is not definable in terms of hydrogen and oxygen atoms, at least not if we consider the 1-intension of our water-concept; the microphysical structure of water was an *a posteriori* discovery, and could not be known merely by inspecting our concepts. (Of course, the 2-intension of water is H_2O – but we did not know this until we discovered that H_2O instantiates the watery properties.) Yet if Chalmers & Jackson are right, then the facts about water are scrutable from the metaphysically basic facts of microphysics. So the question on which everything turns is whether or not scrutability and definability are one and the same thing. Block & Stalnaker say they are; Chalmers & Jackson say they are not.

Chalmers & Jackson are right. This is because definability and scrutability work in different directions. Definability implies an *a priori* entailment from the B-facts to the A-facts; scrutability implies an *a priori* entailment from the A-facts to the B-facts.

Definability means that the higher-order concepts can be analysed in terms of the lower-order concepts. So, if definability is generally true, then we will be able to take any B-fact and transform it into a set of A-facts by an *a priori* translation. If definability is generally true, then for all B-facts there will be A-facts such that it is *a priori* that: if B, then A. But, as we know, this is not always true. It is not *a priori* that if there is water, then there is H_2O. In fact, it is rarely true: whenever a higher-order phenomenon is multiply realisable by different lower-order facts, then there will be a failure of definability.

But scrutability does not require the B-facts to be definable in terms of the A-facts, because it is not about mapping the higher-order facts onto lower-order ones. Scrutability needs the translation to go the other way: we need to map the lower-order facts onto the higher order ones. If the world is scrutable, then for all B-facts there will be A-facts such that it is *a priori* that: if A, then B. Scrutability means that it is *a priori* that if there is H_2O, then there is water.

So it is possible that the world is scrutable, even if higher-order concepts are not generally definable in terms of lower-order ones. When Block & Stalnaker assert that the water facts are not scrutable from the H_2O facts, because in order to deduce them we would need the additional, *a posteriori* premise that water is H_2O, they are guilty of confusing scrutability with definability. Although it is *a posteriori* that the watery properties are instantiated by H_2O, it could still be *a priori* that H_2O instantiates watery properties. In terms of Block & Stalnaker's argument, this means that premise (1) is false. Therefore, even if premise (2) is granted, (3) does not follow.

So scrutability does not require definability. The fact that macrophysical concepts are not generally definable in microphysical terms does not in itself mean that there is no *a priori* entailment from microphysical to macrophysical. Therefore Block & Stalnaker's argument, as it stands, does not succeed. They have not offered a compelling reason to think that the relationship between microphysics and the macrophysical realm is one of metaphysical entailment without *a priori* entailment.

I will end this section with something of a speculative detour. Up to this point, I have assumed (along with Block & Stalnaker and Chalmers & Jackson) that there is a metaphysical entailment from P to M (where M is defined as the macroscopic facts minus consciousness). But what if this is wrong? It is worth considering the possibility that M is not scrutable on the basis of P, and what this would mean.

If M is scrutable on the basis of P, this just means that an ideal observer with complete microphysical information about the world could deduce all the macrophysical facts, without needing any additional *a posteriori* information. We can imagine such an observer having a translation manual that maps the P-facts on to the M-facts. Let us represent the translation manual with the symbol *V*.[18] Then V will consist of statements of the form {if p, then m}, where *p* represents some subset of the total microphysical facts, and *m* represents some corresponding subset of the macroscopic facts (minus consciousness). The fundamental question is whether or not the translation manual, V, is itself *a priori*. If it is, then M will be scrutable on the basis of P; if it is not, then M is not scrutable on the basis of P alone.

I think that in many cases, it is very plausible that the 'if p, then m' statements contained in V will in fact be *a priori*. For example, although it is *a posteriori* that the watery properties are instantiated by H_2O, it is very plausibly *a priori* that H_2O instantiates watery properties – and similarly for the relationship between many macrophysical phenomena and their microphysical realisers. As I have argued in this section, facts about biological life and weather systems are realisable by multiple different distributions of microphysical facts – and indeed, by multiple versions of fundamental physics. So there is no hope of a conceptual analysis of table and chair facts in terms of the microphysics of the actual world. But it is nonetheless very plausible that, given the actual microphysical facts, the biological and meteorological facts follow *a priori*. The biological and meteorological facts are not analysable in microphysical terms; but they are analysable in functional terms, and the microphysical facts realise those functions. In general, for those elements of V that are *a priori*, the corresponding elements of M will be scrutable on

[18] T and M are already taken; I think of the translation manual as a matrix of *vectors* mapping elements of P on to elements of M.

the basis of the relevant elements of P, and there will be no question of metaphysical supervenience without *a priori* entailment.

But are there some parts of V – some statements of the form {if p, then m} – that are *a posteriori*? Are there some elements of V that could conceivably have been otherwise? I think it is plausible that there are. This situation would arise if (and it is, to be fair, a big 'if') there are some macrophysical properties that are essentially connected to phenomenal consciousness. The idea is that some elements of the macrophysical world, though not themselves conscious (there is nothing it is like to be them), are nonetheless partly constituted by how they are perceived by conscious beings. This might seem like a form of anti-realism or idealism – and it is, but only up to a point. The fact remains that P is entirely objective and independent of our cognition. However, this hypothesis would mean that the ordinary macroscopic world, M, is not wholly mind-independent. Rather, M is the product of the interaction between our cognition and the objectively real world of fundamental physics. So the picture is – up to a point – a Kantian one: the ordinary world is a product of the interaction between our cognition and a given element. The major difference between this view and the Kantian one is that in this case the given element does not consist of mental items ('intuitions', in Kant's terminology), but fundamental physics itself. We might call this position *semi-antirealism*.

If some form of semi-antirealism is true, then the corresponding {if p, then m} statements will not be *a priori*. An ideal observer would not be able to deduce the macroscopic facts solely on the basis of P – because the M-facts, in this case, are not fully analysable in terms of functions that can be realised by P-facts. In order to translate from P to M, the ideal observer would also need to know the phenomenal facts relating to how P is perceived by inhabitants of the world. In other words, if some form of semi-antirealism is true, then the translation manual V will not be *a priori*, but will be a subset of C, where C represents the totality of (*a posteriori*) facts about mental representation. And C in turn will depend on the facts about consciousness, Q.

So there is a plausible scenario in which the macrophysical facts are not wholly scrutable from the microphysical ones. This would be the case if some macrophysical properties are partly constituted by their relationship to phenomenal consciousness, assuming that phenomenal consciousness itself is not scrutable from microphysics. Does this help Block & Stalnaker's argument? No. The problem (from their point of view) is that in this scenario, we would also get a failure of *metaphysical* entailment from P to M. If the semi-antirealist view is correct, then we cannot fully isolate the macrophysical facts from the phenomenal facts. In other words, it would be impossible to strictly define M as the macroscopic facts minus consciousness. But in this case, consciousness would also be an essential part of the metaphysical supervenience base of M. The macroscopic facts will supervene, both *a priori* and metaphysically, on the *conjunction* of microphysics with the cognitive scheme with which rational beings – whether God-like conceivers, or us limited, embodied humans – interpret the world. In this case, we would have the following:

$$P + C \Rightarrow M$$

In summary, if we reject the semi-antirealist hypothesis, then it is highly plausible that M is fully analysable in terms of functions that are realisable by P, and that M is therefore scrutable from P. In this case, premise (3) in Block & Stalnaker's argument will be false, and the argument will fail. If, on the other hand, we accept some version of the semi-antirealist hypothesis, then *a priori* entailment from P to M will fail – but so will metaphysical entailment. In that case, premise (4) will be false, and the argument will still fail. Either way, Block & Stalnaker's argument fails, and there is no reason to think that there is metaphysical supervenience without *a priori* entailment.

3.4 - Undecidable Propositions

An undecidable proposition is one that is conceivably true, but also conceivably false, and whose truth or falsity is not a contingent matter. So it seems that they are neither *a priori* true or false, nor contingently true or false. This presents modal rationalism with a dilemma: either they do not have determinate truth-values, in which case we must abandon the logical Law of Excluded Middle; or their truth-value is a matter of strong *a posteriori* necessity.

It seems plausible that there are some mathematical propositions for which there is no proof either that they are true or that they are false. This is not just a matter of our cognitive limitations: even an ideal conceiver would be unable to know *a priori* whether they are true or false. But their truth or falsehood cannot be a contingent matter, since it is absurd to suppose that they could be true in some possible worlds but not in others. They are, after all, mathematical propositions, and it seems hard to make sense of the idea that contingent goings-on could make them true in world A, whilst different contingent goings-on made them false in world B. So we have a dilemma: either such propositions are neither true nor false, in which case we must reject the Law of Excluded Middle; or their truth-value is a strong *a posteriori* necessity.

Chalmers (2010) concedes that:

> Perhaps the most challenging cases for [modal rationalism] are mathematical truths *M* such that *M* is true (and therefore necessarily true and 1-necessarily true) but not knowable, and so not knowable a priori. [p174, author's italics]

The Solution: Chalmers suggests two possible escape routes. The first is that, in his words (p174): 'unprovability within a given system does not entail nonapriority.' So, although it may be entailed by certain mathematical theorems that there are unprovable but true mathematical propositions, this is harmless for modal rationalism just so long as the truths in question are merely unprovable *within a given* system of mathematics. If they are nevertheless knowable *a priori* to an idealised intellect that is not restricted to the axioms of the system in question.

Maybe this approach will be able to deal with all the problem cases. But what if it is not? For the sake of argument, I will suppose that there are indeed mathematical propositions for which no *a priori* proof or disproof exists in any system, even for an idealised intellect. The second possibility is that such mathematical statements are not determinately true or false. But how could this option be embraced without rejecting the law of excluded middle?

My proposed solution, which it is only possible to outline within the scope of this thesis, requires us to draw a distinction between mere *prima facie* propositions and ideal (that is, *real*) propositions. This is an extension of the distinction between mere *prima facie* and ideal conceivability. Consider, for example, whether there is a proposition that the square root of 2 is rational. At one level, it is obvious that there is such a proposition, and a little mathematics shows that it is necessarily false. There is no possible world in which root 2 is rational – but it would be absurd to suggest that the proposition that root 2 is rational is therefore meaningless, or that there is no such proposition. It is only because we can make sense of the proposition that we can prove that it is false. If necessarily false propositions automatically collapsed into meaninglessness, then there would be no such thing as mathematics.

But there is another sense in which there is no proposition that root 2 is rational: to put it simply, there is nothing to propose. Whilst it is *prima facie* conceivable that root 2 is rational, it is not ideally conceivable. But this just means that there is literally no ideally conceivable scenario in which root 2 is rational. The important point is that a merely *prima facie* conceivable scenario is not really a conceivable scenario at all: it is the illusion that there exists a conceivable scenario. But a proposition is a proposal that a particular scenario obtains. In this case, there is no scenario to be proposed – there is literally no way things could conceivably be such that root 2 is rational – and therefore in that sense there is no proposition.

So we must distinguish between mere *prima facie* and ideal propositions: a mere *prima facie* proposition is the apparent proposition that a merely *prima facie* conceivable scenario obtains; and an ideal proposition is the proposition that an ideally conceivable scenario obtains. And there is a sense in which a mere *prima facie* proposition is not really a proposition at all: it is illusory in the sense that it does not really propose a scenario, because there is no real scenario for it to propose.

But this does not mean that mere *prima facie* propositions – and the *prima facie* scenarios that they appear to propose – are meaningless or vacuous. On the contrary, they contain real semantic components with real cognitive value. The *prima facie* proposition that root 2 is rational, for example, contains the real concepts of the number 2, and the notion of a square root, and of rationality. It is precisely because mere *prima facie* propositions contain real semantic values that we are able to prove that they are necessarily false. They are necessarily false precisely because their semantic components do not add up to a coherent whole: there is literally no sum of the parts. If modal rationalism is right, then this is precisely what it means to state that a proposition is necessarily false: not that it proposes a scenario that is prohibited (by necessity) from obtaining, but rather that it proposes no scenario at all. It appears, at face value, that their components add up to a coherent whole; but this is an illusion, and in reality there is no scenario – no way things could conceivably be – that is being expressed.

Now that I have drawn a distinction between mere *prima facie* and ideal propositions, the next step is to recognise that there are two distinct ways in which a merely *prima facie* proposition can fail to be an ideal proposition. The first, as I have just outlined, is if it is incoherent. Like the *prima facie* proposition that root 2 is rational, it fails to present a coherent scenario, and therefore it is necessarily false. But I suggest that there is a second way: this occurs where a proposition appears to adequately define its truth-conditions, but in fact it fails to do so. I propose that this is what is going on with undecidable mathematical propositions. Such propositions are not *a priori* true or false; nor are they contingently true or false; nor is their truth value determined by an *a posteriori* necessity. Therefore they do not have a determinate truth-value at all. This looks like a direct flouting of the Law of Excluded Middle, but I think it is not. The reason that such undecidable propositions lack a truth-value is that their *truth-conditions* are not properly defined, and therefore they are not real propositions (and do not count as exceptions to the Law of Excluded Middle). After all, if a proposition has properly defined truth-conditions, then these will either obtain or they will not. But the puzzling thing about undecidable mathematical propositions is precisely that there is nothing that could possibly render them either true or false: no contingent facts about the world, nor any logical necessity, nor any *a posteriori* necessity (if I am right and there is no such thing). But if there is nothing that could possibly render them true or false, then their truth-conditions are not properly defined, and they are not real propositions. This does not mean that they are meaningless or empty. On the contrary, they contain real semantic components. Ordinary *prima facie* propositions fail to combine their semantic components in a way that yields a coherent whole; similarly, undecidable mathematical propositions fail to combine them in a way that fully determines a set of truth-conditions.

3.5 - Necessary Objects

Another apparent class of strong *a posteriori* necessity concerns the existence of necessary objects. The problem for modal rationalism is that both the following seem to be true:

a) There are necessary objects – that is, objects that exist in all possible worlds.

b) Positive existential claims cannot be known *a priori.*

But if both (a) and (b) are true, there must be strong *a posteriori* necessities relating to the existence of these necessary objects. Are both (a) and (b) true, though?

The most obvious example of a supposedly necessary being is God. It is probably fair to say that the majority of philosophers of religion regard the existence of God as a matter of metaphysical necessity, but not one that can be settled *a priori* either way. That is, the majority of philosophers of religion endorse the proposition that *if* God exists, *then* God exists in all possible worlds – but they will often reject *a priori* arguments for the existence of God, such as the various versions of the ontological argument. Now this position is clearly not consistent with modal rationalism. Modal rationalism presents us with the following dilemma: either it is *a priori* that God exists, and therefore that some version of ontological argument is successful; or it is *a priori* that God does not exist, and there is ultimately no coherent concept of God as a necessary being. I think that is exactly right, and that one or other side of this dilemma must be true – but I will not discuss which in this thesis.

A less controversial reason to accept (a) is that many necessary truths presuppose the existence of certain objects. The obvious example is mathematics, which undoubtedly yields necessary truths, and whose framework is generally thought to presuppose the existence of certain types of abstract objects – numbers, sets, relations, and so forth. For the sake of argument, I will assume that there are at least some necessary objects. So the question for modal rationalism is, how is this possible? How can the existence of anything be *a priori*?

The central idea is that some objects (in particular, certain kinds of abstract object) are such that their existence is wholly constituted by their mere possibility. If they are possible, then there is no further meaningful question about whether or not they exist. And, if modal monism is true, then the mere conceivability of any object constitutes its metaphysical possibility.[19] Therefore the ideal conceivability of such abstract objects is sufficient to constitute their existence. Finally, if something is metaphysically possible, then it is true in all metaphysically possible worlds that it is metaphysically possible. So there are certain abstracta whose ideal conceivability guarantees that they exist in all possible worlds – they are necessary objects, and their existence can be known *a priori*.

[19] One might object that it is question-begging of me to assume the truth of modal rationalism in an argument against an alleged strong *a posteriori* necessity. But I do not think this would be fair. What I am arguing here is that if modal monism is true, then the existence of necessary objects is no counter-example to modal monism; so it is fair to assume the truth of modal monism in my argument.

We can formalise such *a priori* knowledge of the existence of necessary objects by means of the following argument, in which O represents some appropriate abstract object, which may be a number, a set, a relation, a possible world, or other such:

i) It is ideally conceivable that object O exists. [*A priori* definition]

ii) Ideal conceivability entails metaphysical possibility. [Modal monism – which I will argue in the next chapter is an *a priori* truth]

Therefore:

iii) It is metaphysically possible that object O exists.

Therefore:

iv) It is true in all metaphysically possible worlds that object O is metaphysically possible. [Follows *a priori* from (iii) and the definition of metaphysical possibility as unrestricted possibility]

iv) It is an *a priori* truth that: if it is metaphysically possible that O exists, then O exists.

Therefore:

v) Object O exists in all possible worlds.

Clearly it is premise (iv) that is doing most of the work here. My contention is that there are *a priori* truths of this form, such that the mere possibility of certain abstract objects constitutes their actual existence. The important thing is that the formal structure of premise (iv) does not quantify over some object O, since this would obviously beg the question of whether or not there are any such objects, as formulas (e) and (f) do. It simply asserts the *a priori* truth that if it is possible that O exists, then O exists. I think it is highly plausible that this is precisely true with respect to such possible abstract objects as numbers, relations, possible worlds, and so forth. After all, given that it is possible that the number nine exists, what more is required for it to actually exist? It seems to me inconceivable that there is a world in which the number nine, being possible, nonetheless happens not to exist. So if this is right, then we have an explanation of how we can have *a priori* knowledge of the existence of necessary objects.

Conclusion

In this chapter, I have considered some of the more important alleged cases of strong *a posteriori* necessity: demonstratives and indexicals; essential properties of individuals; metaphysical supervenience; undecidable propositions; and necessary objects. Each of these cases seems to present us with a scenario that is only knowable *a posteriori*, but which is metaphysically necessary: that *this* is H_2O; that David's father is Owen; that the microphysical structure of the world determines its macroscopic structure; that some mathematical proposition P is true; that numbers exist. But I have argued that in each case the appearance is illusory. None of the supposed examples generates a scenario, S, such that S is ideally conceivable but metaphysically impossible. However, the fact that none of the cases I have considered in this chapter are successful does not mean that there are no strong necessities at all. This much stronger claim is what I will argue in the next chapter: that there are no strong *a posteriori* necessities at all, because they are impossible.

Chapter 4: The Deep Incoherence of Modal Dualism

Introduction

We have seen that ideal conceivability entails metaphysical possibility if and only if there are no genuine *a posteriori* necessities. In the last chapter, I considered several alleged examples of *a posteriori* necessity, and argued that none of these are convincing candidates. Yet it is one thing to argue that – so far – we have not been presented with a single case of an unambiguous, genuine *a posteriori* necessity; it is quite another thing to claim that there are none. That is what I will argue in this chapter. In fact, I will argue for a stronger thesis: that there *cannot* be any *a posteriori* necessities. I will argue that epistemic possibility and metaphysical possibility are one and the same, and that the attempt to separate them collapses into incoherence. If this is true, then the ideal conceivability of a scenario will entail that there is a corresponding metaphysically possible world. The reverse is also true: every metaphysically possible world must be epistemically possible. This means that, just as there cannot be any *a posteriori* necessities, so there cannot be any *a priori* contingent truths.

In Section 4.1, I will explain in more detail the contrast between modal monism and modal dualism, setting out the version of modal monism that I advocate. In Sections 4.2, I outline the standard arguments for modal monism, and explain why these do not address the central issue; then in Section 4.3 I set out my central argument for modal monism, which is that there is no coherent way to formulate modal dualism. In Section 4.4, I address an apparent problem with modal monism, which is that it seems to rule out counter-possible reasoning. Finally, in Section 4.5, I address supposed cases of the contingent *a priori* and show that they are vulnerable to the same deflationary two-dimensional analysis as the Kripke cases, and therefore do not count against modal monism.

4.1 - Modal Monism v Modal Dualism

We can define two fundamental and notions of possibility (and, correspondingly, of necessity). The first is *metaphysical* possibility. Metaphysical possibility is defined as the most fundamental type real possibility there is. It is, by definition, unrestricted: metaphysical possibility is not defined by placing some additional restriction on some more basic realm of possible worlds, because there is no more basic realm of worlds. Metaphysical possibility is possibility *simpliciter*.

The second is *epistemic* possibility. Something is epistemically possible from my point of view if, for all I know, it could be the case. Similarly, something is epistemically necessary from my point of view if, given what I know, it must be so. We can then expand this notion to cover the limiting case of an ideal rational conceiver, who is not allowed to use any *a posteriori* premises in their reasoning. We can then ask: what could be true, given what such a conceiver would know? And what must be true? In this way, we can define a notion of ideal epistemic possibility.[20]The notion of ideal epistemic possibility is therefore intimately connected to that of conceivability. If a scenario is ideally conceivable, then there is a corresponding epistemically possible world, and vice versa; and if a scenario is not epistemically possible, then it is not really ideally conceivable, and vice versa.

There are, of course, other varieties of possibility and necessity that describe some notion of what could or must be the case: causal necessity, natural necessity, and so forth. These can be understood as restrictions placed on a more fundamental space of possible worlds; none of them seem, at face value at least, to be fundamental. For the sake of argument, I will set all these other notions aside, and focus only on the two notions of metaphysical and epistemic modality. The question is: what is the relationship between these two fundamental types of possibility? Are they really distinct? Or are they ultimately one and the same thing?

So we have two opposing views of the space of possible worlds. Modal monism (Chalmers 2010, Schroeter 2012, 2.3) is the view that metaphysical modality and ideal epistemic modality are fundamentally one and the same. Thus every world that is epistemically possible is metaphysically possible, and vice versa. According to this view, the space of possible worlds is determined by what is ideally conceivable, and there is no distinct metaphysical modality. This is not to deny that there is such a thing as metaphysical necessity – it is just to deny that it is distinct from epistemic necessity.

[20] It is important to bear in mind that when I refer to epistemic modality, unless otherwise specified, I am always talking about this idealised, limiting-case notion of what is possible or necessary from the perspective of an ideal rational conceiver, who is not permitted to use any *a posteriori* premises in their reasoning..

Modal dualism (e.g. Edgington 2004) in contrast is the view that there are two fundamental types of modality: there is epistemic possibility, and, distinct from this, there is also metaphysical possibility. There is then a further question about the extent to which these categories overlap: there may be some epistemically possible worlds that are not metaphysically possible – and perhaps even some metaphysically possible worlds that are not epistemically possible. As Edgington (2004) puts it:

> […] there are two independent families of modal notions, metaphysical and epistemic, neither stronger than the other. Metaphysical possibility is constrained by the laws of nature. Logical validity […] is best understood in terms of epistemic necessity. [p1]

Thus, according to modal dualism, there is an epistemic notion of modality, which is connected to logical necessity (and thus to conceivability and the *a priori*); and there is a metaphysical notion, which is grounded in some other realm of facts, which in Edgington's view consists of the laws of nature. I will not address the specific question of whether it is plausible that metaphysical modalities are grounded in the laws of nature. The important point here is what they are not: they are not fundamentally epistemic, and not fundamentally reducible to matters of conceivability and the *a priori*.

It is easy to see how this debate relates to the question of whether there are strong *a posteriori* necessities. If modal monism is true, then every ideally conceivable scenario, in constituting an epistemically possible world, must thereby constitute a metaphysically possible world. So there cannot be any strong *a posteriori* necessities. (By the same logic in reverse, there cannot be any contingent *a priori* truths either.) Hence modal monism entails modal rationalism. But if modal dualism is true, this at least raises the *prima facie* possibility that there are epistemically possible worlds that are not metaphysically possible – that is, of strong *a posteriori* necessities.

A related issue is whether or not the space of metaphysically possible worlds varies depending on which world is considered as actual. This is an issue on which modal monism and modal rationalism fundamentally differ: modal dualism entails that, in some cases, the metaphysical modalities are relative to which epistemically possible world is actual; but modal monism must rule this out.

Now, this does not mean that, according to modal dualism, the metaphysical modalities vary from world to world. Modal dualists will rightly maintain that it is metaphysically necessary that water is H_2O, and that twin-water is XYZ; and that this is true in all (metaphysically) possible worlds. The point is that, if modal dualism is true, then the metaphysical modalities vary depending on which *epistemically* possible world is considered as actual. Ultimately this is because modal dualism entails that there are *a posteriori* necessities, and one of the features of *a posteriori* necessities is that different epistemically possible worlds, when considered as actual, can yield different spaces of metaphysically possible worlds.

For example, consider the water / H_2O case, and the following thought-experiment: Suppose that, for the last hundred years and more, chemists had been systematically deceiving the world with respect to their knowledge of the microphysical structure of water. All this time, they have claimed to know that water is H_2O – but now it emerges that the evidence was uncertain, doubts were suppressed, and they are in fact far from certain that water really is H_2O. What they have managed to establish is that either water is H_2O, or it is XYZ – but they are not yet certain which. Moreover, they have just (fortuitously) developed the means to confirm one way or the other, and settle the matter once and for all. Now that their deception has been exposed, the leading chemists of the world are sent away to conduct the necessary experiments and determine the true microphysical structure of water, while the world awaits the verdict.

So in this scenario, we have two epistemically possible worlds – in W1, water is H_2O; in W2, it is XYZ – and we have no idea which one we are in, until the chemists deliver their verdict. Moreover, these worlds are not just epistemically possible from our hypothetical perspective; they are ideally epistemically possible (provided the concept 'water' is understood in terms of it primary intension – once we have assigned a secondary intension to it we have, of course, begged the question). What then does the space of metaphysically possible world look like in this scenario? According to modal dualism, we cannot know until the chemists report back with the results of their experiment. If it then turns out that we are actually in W1, then it is metaphysically necessary that water is H_2O. But if we are actually in W2, then it is metaphysically necessary that water is XYZ. The chemists will not just be discovering a chemical fact; they will be discovering a metaphysical one. Thus in this case the space of metaphysically possible worlds will vary, depending on which epistemically possible world turns out to be actual.

So modal dualism entails that, for different epistemically possible worlds considered as actual, there will be distinct spaces of metaphysically possible worlds. But, as I noted above, this is not the same thing as claiming that the metaphysical modalities vary from world to world – clearly they do not. Nor, at face value, is it the same as claiming that it is *metaphysically* possible that the metaphysical modalities might have been other than they are. That is obviously absurd: it would mean that the facts about metaphysical necessity are themselves contingent. What it does mean is that modal dualism entails that it is ideally epistemically possible – it is ideally conceivable – that the metaphysical modalities might have been other than they are.

Hence we see the very different responses that modal monism and modal dualism give to the Kripke cases. According to modal dualism, the Kripke cases show that the metaphysical modalities depend upon the *a posteriori* facts: if the *a posteriori* facts had been different, the metaphysical modalities would have been different (and it is at least epistemically possible that they could have been different, even if not metaphysically possible). But, according to modal monism, this appearance is misleading. It is not the space of metaphysically possible worlds itself that varies according to the *a posteriori* facts, but the secondary intensions of (some of) our singular terms. This means that the way in which we *describe* the space of possible worlds will depend upon the facts contained in the actual world; but the underlying modalities themselves are unchanging.

The version of modal monism that I am advocating is similar in most respects to that developed by Chalmers. However, there is one respect in which I diverge from his formulation. Chalmers sometimes equates the 1-modalities with epistemic modalities, and – crucially – the 2-modalities with metaphysical modalities. This is evident in his discussion of the structure of the two-dimensional argument (2010, p149), and is most explicitly stated as one of the central theses of his version of modal rationalism:

> (T4) A sentence token S is metaphysically necessary iff the secondary intension of S is true at all worlds. [2010, p546]

At face value it seems plausible to equate the 1-modalities with the epistemic modalities, and the 2-modalities with the metaphysical. But, on reflection, this is inconsistent with a thorough commitment to modal monism. After all, the space of 2-modalities – represented by the horizontal rows of the two-dimensional tables – *does* vary depending on which world is considered as actual – that is, according to the *a posteriori* facts. The whole point of modal monism is that the space of metaphysically possible worlds is identical to the unchanging, underlying space of epistemically possible worlds. Therefore, we want the metaphysical modalities to remain as constant as the epistemic ones – and not to track the 2-intensions, as they in turn track the *a posteriori* (and indeed contingent) facts that vary from world to world. Thus it can hardly be right to equate 2-modality with metaphysical modality.[21]

The notation that Chalmers employs to express two-dimensional semantics (and which I also adopt) does lend itself to the improper identification of 2-possibility with metaphysical possibility. Once we are in the habit of talking about 1-possibility and 2-possibility, and given that we are trying to explain the relationship between epistemic necessity and metaphysical necessity, it seems natural to draw the false equivalence. But, properly speaking, 1-possibility and 2-possibility are not different types of possibility; nor are 1-conceivability and 2-conceivability different kinds of conceivability. The only difference lies in the nature of the underlying concept that we are invited to conceive, or whose possibility we are considering: that is, whether it is a primary or a secondary intension. 1-possibility is shorthand for the possibility of a scenario when interpreted according to the primary intensions of the concepts involved; 2-possibility is shorthand for the possibility of a scenario when interpreted according to the secondary intensions of the concepts involved. But we are not really dealing with two kinds of possibility, any more than 1-conceivability and 2-conceivability are really two kinds of conceivability. Possibility is just possibility, and conceivability is just conceivability.

[21] Of course, 2-possibility is an instance of metaphysical possibility, namely metaphysical possibility with respect to a 2-intension. But, by the same token, 1-possibility is just metaphysical possibility with respect to a 1-intension. The point here is that it is a mistake to equate metaphysical possibility with 2-possibility. This raises the question of why, if 2-possibility is not identical to metaphysical possibility, it is nevertheless specifically the 2-possibility of zombies that is held to be inconsistent with physicalism. I will answer this question in Chapter 5.

We can see all of this at work if we look again at the two-dimensional table for ‘water is H_2O’:

‘Water is H_2O’		World considered as **Counterfactual**	
		W1	**W2**
World considered as **Actual**	**W1**	T	T
	W2	F	F

In this table, the 1-intension interpretation corresponds to the diagonal from top left to bottom right. It is thus 1-possible that water is H_2O iff there is a cell in this diagonal which contains the value ‘True’. The 2-intensions correspond to the horizontal rows. There is therefore a different 2-intension depending on which world is considered as actual. This is exactly as we would expect. But we also find that if we take W1 as actual, then it is 2-necessary that water is H_2O; if we take W2 as actual, then it is 2-necessary that water is not H_2O. But, as I argued above, we do not (if we are modal monists) want the space of metaphysically possible worlds to depend upon the *a posteriori* facts of the actual world. We do not want there to be one space of metaphysically possible worlds in the event that we are in W1, and another space if it turns out that we are in W2, as this would lead to modal dualism. So, if modal monism is true, the horizontal rows do not each represent a different space of metaphysically possible worlds; rather we should think of each row as representing a *way of describing* the space of 1-conceivable worlds (which are the same as the metaphysically possible worlds) from the point of view of a given world which is considered as actual.

4.2 - Modal Dualism and the Problem of Brute Necessities

Now that I have set out the contrast between modal monism and modal dualism, we can turn to the question of which is true. Chalmers (2010) lists several reasons for thinking that there are no strong *a posteriori* necessities:

> […] (i) strong necessities cannot be supported by analogy with other a posteriori necessities; (ii) they involve a far more radical sort of a posteriori necessity than Kripke’s, requiring a distinction between conceptual and metaphysical possibility at the level of worlds; (iii) they lead to an ad hoc proliferation

> of modalities; (iv) they raise deep questions of coherence; (v) strong necessities will be brute and inexplicable […] [p170]

What should we make of these reasons? Point (i) is correct – after all, the whole point of strong necessities is that, unlike the standard Kripkean cases, they are not vulnerable to the deflationary two-dimensional analysis. It is for this reason that point (ii) is also correct, as I have set out in this section. But this does not constitute an argument against strong necessities – it is merely an analysis of what they involve, and what is it stake.

The force of point (iii) – the charge that modal dualism leads to an ad hoc proliferation of modalities – depends on whether or not there are strong necessities. Certainly, modal dualism represents a proliferation of fundamental modalities when compared to modal monism; and this arguably constitutes a reason to prefer modal monism, other things being equal. But the question is whether or not other things are equal – that is, whether or not both theories are equally able to explain the modal data. And this boils down to the question of whether or not there are strong *a posteriori* necessities. The *ad hocness* of having multiple fundamental modalities will only count as a reason against modal dualism if such proliferation is unnecessary – that is, if the resources of modal monism are sufficient to explain the modal data. So, whilst I agree with Chalmers's point (iii), I do not think that it provides independent grounds for preferring modal monism to modal dualism.

Points (iv) and (v) do offer substantial, independent grounds to prefer modal monism to modal dualism, and I will expand on them in this section and the next. In my view, the decisive consideration is Chalmers's suggestion in point (iv) that modal dualism is deeply incoherent; that is what I will argue in the next section. In the remainder of this section, I will expand on point (v), and argue that in fact the problem of brute necessities is more significant than often realised.

Let us suppose that some strong *a posteriori* necessities do exist – what might they be like? Advocates of strong necessities usually imagine that they will embody grand metaphysical truths, such as the impossibility of zombies, or at least deep truths of nature, such as the essences of things. But there is no good reason to suppose that strong *a posteriori* necessities, if they are possible, should not crop up in all sorts of unexpected places.

Consider some proposition, P, that is ideally epistemically possible, but not epistemically necessary. An ideal rational conceiver, on the basis of rational reflection alone, could not know whether or not P is true. Now further suppose that P is also epistemically possible, but not epistemically necessary, relative to me at the present time. Just as ideal rational reflection is not sufficient to determine the truth of falsehood of P, so my present stock of knowledge is insufficient. For all I know now, P could be true, or it could be false. But what about the modal status of P – can I know anything about that? In particular, can I rule out the existence of a strong *a posteriori* necessity in respect of P?

I say that, if strong necessities are possible, then for any P that is epistemically possible, there is no way to rule out the existence of a strong necessity in respect of P. So, for all P, if P itself is epistemically possible, then it must be epistemically possible that P is metaphysically necessary; and if P is not epistemically necessary, then it will be epistemically possible that P is metaphysically impossible.

But – and here is the problem – P could be *anything*. Or at least, it could be anything that is epistemically possible to me. It could (epistemically) be that P represents the impossibility of zombies, or that water is H_2O. But it could equally be that P is something that we would intuitively think ought to be a contingent matter, such as the outcome of a coin-toss, or the occurrence of a natural event. P might even be such as to over-ride our normal expectations regarding the future behaviour of the world: P might, for example, be the proposition that tomorrow the laws of nature will change in some specified way. For all of these values of P – and indeed, for any others that are epistemically possible – it will be epistemically possible that P is metaphysically necessary.

Now, advocates of strong necessities might try to resist this conclusion by claiming that strong necessities must conform to certain principles, and are therefore limited to special cases, such as identities and the impossibility of zombies. This could allow for the possibility of some strong necessities, whilst preventing their wild proliferation.

But I do not think this limitation strategy will work. For a start, it is not enough to appeal to principles that will deliver only weak necessities. It may well be that my existing stock of knowledge and my understanding of how my concepts work give me reason to think that there is a metaphysical necessity in relation to P, even if I do not know whether P is true or false.

Suppose that I do not know the microphysical structure of water, and P is the proposition that water is H_2O. If my water-concept rigidly designates whatever microphysical substance plays the relevant watery role in the actual world, then I can know that P is either metaphysically necessary, or metaphysically impossible, even though I do not know which. But of course, in this case we will only have a weak necessity: P will have a secondary intension that is either metaphysically necessary or metaphysically impossible, but the primary intension of P will be contingent. But weak necessities are not sufficient for modal dualism. But the limitation strategy requires principles that yield strong necessities. Are there any such?

Let us suppose that there are. Let us imagine that there are some principles (I know not what) that inform us that some P is the sort of proposition that, whether true or false, will be a strong *a posteriori* necessity. Such principles would effectively say: 'Look to these types of propositions for strong necessities'. Perhaps, for the sake of argument, there are some such principles – and so, for some types of P, I can know in advance that whether P is true or false, there will be a strong *a posteriori* necessity with respect to P.

However, even this is not enough for the limitation strategy to work. Even if there are some cases where I can know in advance that there is a strong necessity with respect to P, because P belongs to the relevant class of propositions, the limitation strategy requires the stronger claim that if P is *not* in the relevant class, then I can *rule out* the existence of a metaphysical necessity in respect of P. Suppose that some proposition P lies outside the class of propositions in respect of which, by some hypothetical set of principles, we would expect to find strong necessities. Does this mean we can know for sure that there is *not* a strong necessity in respect of P? No, because if it is conceivable that there are brute *a posteriori* necessities, then it will be conceivable that there is a brute, *sui generis*, inexplicable necessity that is unique to P.

The only conceivable way to rule out the possibility of a brute, *sui generis* strong necessity will be if there is an *a priori* principle that restricts strong necessities to those in the privileged class of propositions, those to which the relevant principles apply – a limitation strategy. For example, an advocate of strong necessities might propose that it is an *a priori* principle that:

> If object *a* has super-explanatory property *b*, then necessarily *a* has *b*.

A super-explanatory property (cf Godman, Mallozzi & Papineau, forthcoming) is a relatively more fundamental property that causally (or otherwise) explains a large number of less fundamental, observable properties of an object. And, according to this strategy, it is *a priori* that super-explanatory properties are metaphysically necessary. We can then substitute actual instances of objects and their (*a posteriori*) super-explanatory properties to generate strong *a posteriori* necessities. For example, having the chemical composition H_2O is super-explanatory with respect to the observable properties of water; being fathered by Owen is (partially) super-explanatory of the observable physical characteristics of David; and so forth. But the fact that it comes out of taps is not super-explanatory of water's properties; and being born in Como is not super-explanatory of David's properties – hence why these things are not metaphysically necessary.

But I do not think this limitation strategy works. For a start, a successful strategy requires an *a priori* principle that strong necessities are *only* to be found amongst a certain class of truths. But, even by its own lights, the super-explanatory principle does not restrict strong necessities to those involving super-explanatory properties – it just entails that super-explanatory properties yield strong necessities. Furthermore, I do not think that the super-explanatory principle is plausibly *a priori* as it stands. In fact, it is not even always true. For example, consider the generic type *watery stuff* (which includes water and twin water). Having the microphysical structure H_2O is super-explanatory of the cluster of properties that constitute a substance being *watery stuff*. But it is not necessary that watery stuff has the microphysical structure H_2O – there are other possible microphysical structures that could play the super-explanatory role. Of course, the super-explanatory principle will be *a priori* if we amend it so that if object *a necessarily* has super-explanatory property *b*, then necessarily *a* has *b*. But this is obviously question-begging.

What is plausibly *a priori* is that if object *a* has super-explanatory property *b*, and we intend that our concept of object *a* should track its super-explanatory properties, then necessarily object *a* has property *b*. I think this is right – but this principle will only yield weak necessities. For example, we intend that my concept of a human individual should track the super-explanatory properties, whatever they turn out to be, of the observable physical characteristics of that individual. Therefore, if it turns out that *a* was fathered by *b*, then no hypothetical individual will count as being *a* unless that individual was fathered by *b* – even if the hypothetical individual exactly resembles *a* in respect of all the observable physical properties.

But this is just like the standard Kripkean cases. We have to distinguish between two levels: there is the *a posteriori* fact that *b* is super-explanatory with respect to the observable characteristics of *a*; and there is the conceptual fact that, given this *a posteriori* fact, no possible object will count as an instance of *a* unless it has super-explanatory property *b*.

In summary, for the limitation strategy to work, it is not enough to put forward a set of principles that will only yield weak necessities – the principles in question must yield strong necessities. Furthermore, even if there were any such principles, it is not enough for them to identify certain cases where we should expect to find strong necessities: they need to rule out the possibility of strong necessities outside of these cases. And this is something that they cannot do, because if strong necessities are possible, there will always be the epistemic possibility of a *sui generis*, unique strong necessity outside the scope of the principles. The only the limitation strategy would succeed would be if there were an *a priori* principle that strong necessities occur in certain cases, and only these cases. The super-explanatory principle appears to offer such a principle, but on closer inspection it does not succeed. So the limitation strategy will not work.

The moral is that, if strong necessities are possible, for any proposition P that is epistemically possible (whether ideally or for a particular subject at a given time), it will be epistemically possible (in the same sense) that P is metaphysically necessary. But this means that anything that is epistemically possible could, for all I know, be metaphysically necessary. If strong necessities are possible, then there will be no reason to think that they are restricted to grand metaphysical principles or deep truths of nature; for all I know, reality could be infested with strong necessities, many of them brute and arbitrary.

However, while the problem of brute necessities is an important issue, I do not believe it is the fundamental problem with modal dualism. Moreover, the debate between modal monism and modal rationalism is not a matter of choosing between rival, equally coherent theories and selecting the one that, on balance, provides the best explanation of the data. Of much greater importance is the problem outlined in Chalmers's point (iv) – that modal dualism raises deep questions of coherence.

In fact, as I will argue in the next section, modal monism is itself an *a priori* truth. This is far from obvious. Nevertheless, I believe it to be true. What is clear, however, is that if true, modal monism must be *a priori*: either it is an *a priori* truth, or it is false. Why so? Because if modal monism is true, then there are no strong *a posteriori* necessities. But what sort of truth would this itself be – that there are no *a posteriori* necessities? It could hardly be a contingent truth that there are no *a posteriori* necessities - if that were so, it would be true in some possible worlds and false in others. But whether or not there are strong *a posteriori* necessities is a truth about the space of possible worlds as a whole, so it cannot be true in some worlds and false in others. Therefore, if there are no strong *a posteriori* necessities, this must itself be a necessary truth. But it cannot itself be an *a posteriori* necessity that there are no *a posteriori* necessities; that would be obviously self-refuting. Therefore, if there are no strong *a posteriori* necessities, then it must be an *a priori* truth that there are no strong *a posteriori* necessities. Hence the dilemma: either strong necessities exist, or it is *a priori* that modal dualism false.

But is this right? Are strong necessities really *inconceivable*? Modal dualism is clearly at least *prima facie* conceivable. Enough people believe in strong *a posteriori* necessities to demonstrate that they are not obviously incoherent. But, though it may not be obvious, that is what I will argue in the next section.

4.3 – The Deep Incoherence of Modal Dualism

The central against modal dualism aims to show that there is, ultimately, no coherent way to formulate it. This argument has three steps: first, that the standard formulation of modal dualism is incoherent; second, that the only even *prima facie* plausible way to reformulate it entails a commitment to the existence of impossible worlds; third, that the notion of impossible worlds required by modal dualism is incoherent. But if modal dualism entails an incoherent commitment, then it must be false; so I shall argue in this section.

For the reasons set out in section 4.2, modal dualism entails that there are at least some strong *a posteriori* necessities. But what does this mean? It means that there is at least one instance in which there is an ideally conceivable scenario that is not metaphysically possible. Now, according to the definition of modal dualism that I set out in the previous section, ideal conceivability constitutes one of the two fundamental types of possibility, namely (ideal) epistemic possibility. So we have an instance of a world that is epistemically possible, but not metaphysically possible.

But now we have a problem, because it is also supposed to be true by definition that metaphysical possibility is the most fundamental, unrestricted notion of possibility that there is. That is, metaphysical possibility is possibility *simpliciter*. Indeed, this is a point on which modal monism and modal dualism are in agreement. But how can metaphysical possibility be unrestricted possibility if there are some worlds that exhibit another type of possibility – epistemic – without being metaphysically possible? This is a straightforward contradiction. We cannot have it both ways. It cannot be the case that metaphysical possibility is possibility *simpliciter*, and that it only applies to some narrow subset of a broader class of worlds that are possible in some other sense.

What this shows is that the standard way of formulating modal dualism is straightforwardly incoherent. It is impossible for all of the following three things to be true: first, that metaphysical possibility is possibility *simpliciter*; second, that ideal conceivability constitutes another fundamental type of possibility, namely epistemic; third, that there are strong *a posteriori* necessities.

So how can we reformulate modal dualism to avoid this contradiction? There is no point in abandoning either the first or third of these claims, since the first is a straightforward and uncontroversial definition, and the third is fundamental to modal dualism. So the only hope is to abandon or reformulate the second claim. So modal dualism, if it is a coherent possibility, must abandon the notion that ideal conceivability constitutes epistemic possibility. This should not come as a surprise. If modal dualism is true, then ideal conceivability does not automatically constitute metaphysical possibility. But, since metaphysical possibility is meant to be the most fundamental type of possibility that there is, it follows that if modal dualism is true, then ideal conceivability does not automatically constitute any type of possibility at all. Modal dualism requires that we can ideally conceive of a world, W, such that W is not possible in *any* sense, including epistemically.

But this presents a problem. When we judge that a world is conceivable, it is natural to think that it therefore represents a *way things could conceivably be*. But now it is already too late: we have slid effortlessly from a world's conceivability to the conclusion that it is at least epistemically possible. So, if modal dualism is a coherent position, then there must be a way of resisting this slide from the ideal conceivability of W to the epistemic possibility of W. But then what are we conceiving of, when we conceive of W, where W is metaphysically impossible? We cannot be conceiving of a way things could be. So, the only recourse for modal dualism is that W is a *way things could not be* – that is, an impossible world. Therefore, modal dualism entails that we (or rather, an ideal conceiver) can ideally conceive of W, such that W is an impossible world. Moreover, the requirement is not merely that we can ideally conceive of W, and that W happens incidentally to be impossible. If that were the case, then our conception of W would be of a way things could conceivably be – and it would be an epistemic possibility, which is the outcome that modal dualism needs to avoid. Rather, the requirement is that the impossibility of W is part of our conception of W: in conceiving of W, we are conceiving of a way things could not be. In conceiving W, we are conceiving of an impossible world.

So, if modal dualism is a coherent position, then we will have to modify the traditional way of formulating it. Rather than conceiving of a broader space of epistemically possible worlds that contains a narrower subset of metaphysically possible worlds, we must picture the space of metaphysically possible worlds – all the ways that things could be – surrounded by a broader space of ways things could *not* be – a space of impossible worlds, some of which, such as W, will be ideally conceivable.

But why is this a problem? This brings us to the question of whether or not there are impossible worlds. There seem to be good arguments both ways: on the one hand, it seems obvious that there are impossible worlds; but on the other hand, it seems clear that impossible worlds cannot exist.

On the one hand, it seems to be a truism that there are ways that things could not be (cf Berto, 2013). The world could not be made of square triangles, for example; it could not be such that the square root of 2 is rational. So, these conceptions clearly represent ways things could not be. And the natural thought is that we can quantify over these *ways that things could not be* in exactly the way that we quantify over *ways that things could be*. I have, in fact, just listed two ways things could not be. And, just as the *ways that things could be* are possible worlds, so the *ways things that could not be* are impossible worlds. Therefore there are impossible worlds.

But on the other hand, there are good reasons to think that there cannot be impossible worlds. Surely, there cannot be – there cannot exist – a thing that is impossible. The notion that there could be an impossible world is based on a misconception of what a possible world really is. It is a misconception to think of a possible world as merely one type of world, existing alongside the impossible ones. A possible world is not a type of world, namely a world which happens to be possible – it is a type of possibility, namely the possibility of a world. If there is no possibility, then there is no world. Therefore, if something isn't a possible world, then it isn't a world at all – and, *a fortiori*, it cannot be an impossible world.

So, we face the dilemma that on the one hand there clearly are impossible worlds – and yet there cannot be. How can we solve this dilemma? The answer, as is often the case, is that we have an ambiguity: there is a sense in which it is perfectly harmless to speak of there being impossible worlds; and there is a sense in which it is deeply incoherent.

What is the difference? Of the two innocent examples of impossible worlds that I gave above, one was at best *prima facie* conceivable (namely a world in which the square root of two is rational) and the other (the word of square triangles) was not even *prima facie* conceivable, being obviously incoherent. In neither case were we presented with an ideally conceivable scenario, subsequently to learn that it is not metaphysically possible. What is really going on here is that we have an (at best) mere *prima facie* conceivable scenario that is not in fact ideally conceivable, and therefore does not correspond to any possible world. So, when we say that a world in which the square root of 2 is rational is an impossible world, what we really mean is that there is no world corresponding to the mere *prima facie* conception that root 2 is rational. Although English grammar forces us to quantify over impossible worlds, the underlying logic does not.

So what we are really doing in the innocent case is quantifying over mere *prima facie* conceptions (or, in the square triangle case, obviously incoherent ones), and denying that there exists a corresponding possible world. (*Prima facie* conceptions themselves exist unproblematically, so there is no problem in quantifying over them). What we are not doing is picking out a world and asserting that it has a property, namely the property of impossibility. When we seemingly make a positive existential claim about impossible worlds, we are really making a negative existential claim about possible worlds.[22]

[22] There is a close analogy, I think, between impossible worlds and the problem of non-existent objects. There is perfectly legitimate sense in which there are non-existent objects, such as Father Christmas, the rational square root of 2, and so on. In these cases, we have a conception (whether ideal or merely *prima facie*) that does

So, there is a perfectly legitimate and coherent sense in which there are impossible worlds: there are mere *prima facie* conceptions, for which no possible world exists. An impossible world is not a world that is prohibited by a law, whether of logic or metaphysics: it is the absence of a possibility, where our imperfect powers of conception make it appear as if there is one.[23]

However, what we cannot coherently do is identify a world, W, such that W exists and has the property of impossibility. Now, it may be objected that I am trading on an ambiguity of scope here. Of course, it would be incoherent to assert that there exists a world that could not possibly exist. But, so the objection goes, that is not what is meant by an impossible world. All that is meant by an impossible world is a world that could not possibly be actualised. And there is nothing obviously incoherent in supposing that such worlds exist. But this objection, as plausible as it sounds, is not correct. There is no ambiguity of scope here, since there is no meaningful distinction between the possibility of a world being actual, and its existence as a possible world. A world that could not possibly be actual does not exist as a possibility, in which case it does not exist as a possible world. But if it doesn't exist as a *possible* world, then it doesn't exist as a world at all, because a world is just a possible way things could be. The mistake is to think of the possible worlds as a subset of the worlds, namely the possible ones, when they are really a subset of the possibilities, namely those that describe the world as a whole.

Nor should we be tempted by the thought that, since (as I outlined in Chapter 2) possible worlds can be regarded as world-sized, maximally-detailed properties (namely the property of

not in fact refer to any object. So we can quite legitimately say that Father Christmas does not exist – or, if we want, that Father Christmas is a non-existent object (the man, that is, as opposed to the concept of Father Christmas or fictional entity, both of which actually exist *qua* concept and fictional entity, respectively). But we must interpret this truth as making a negative existential claim, not a positive one. We are saying that there does not exist an object to match a certain concept of ours, namely that of Father Christmas. What we are not doing is picking out an object and attributing to it the property of non-existence. It would be quite wrong to say that there exists an object that is non-existent. Impossible worlds, in fact, are a special case of non-existent object.

[23] An impossible world is like a stone that is too heavy for God to lift. Can God make such a stone? No – but, as the theologians insist, that is no limitation on God's power. The reason God cannot make such a stone is that the very concept is incoherent. It is not that God could try to make such a stone, only to fail every time – there is nothing there for God to even attempt.

being a certain way), this means that, amongst the uncountable infinity of such properties that are in fact uninstantiated, there may be some that are *uninstantiable* – that is, which represent impossible worlds. What is wrong with this idea? At face value it seems conceivable that there are some such properties that are uninstantiable. But if any such property were uninstantiable, that would mean that there is no possible world that instantiates it. But if there is no possible world that instantiates it, then there is no property. The property just is the possibility of the world being a certain way. The possibility of the property *qua* property is one and the same thing as its possible instantiation by a world. Thus, if there is no possibility of the world being a certain way, then there is no corresponding property.

So it is meaningful to speak of there being impossible worlds, but only so long as this is understood correctly: it is not a positive existential claim about worlds that have the property of impossibility, but rather a negative existential claim, that for mere *prima facie* conceptions, there do not exist corresponding possible worlds. If it is merely *prima facie* conceivable that W* is actual, then it will be merely *prima facie* conceivable that there is a possible world W*; in which case, we can legitimately say that W* is an impossible world. But what is not coherent is to assert that there exists a world W, such that W is not possibly actualised.

This means that there is no coherent way to satisfy the requirements of modal dualism. It is not enough for modal dualism that, in conceiving of an impossible world, we have a mere *prima facie* conception of a way things could be. Modal dualism requires the much stronger claim that we (or rather, an ideal conceiver) can ideally conceive of a way things could not be. In other words, modal dualism requires that we can ideally conceive of W, such that W is not possibly actualised. But we can now see that this is an incoherent requirement. It requires us to conceive of a world that exists *qua* world, but is not possibly actualised; but a world that is not possibly actualised is not a possible world; and a world that is not possible is not a world at all.

But, it may be objected by the modal dualist, ideal conceivability need not be cashed out in terms of worlds at all – possible or otherwise. Even if we cannot ideally conceive of an impossible world, perhaps we can ideally conceive without making any commitment in terms of worlds. So, the argument goes, we can ideally conceive without entailing the existence of an epistemic possibility – and therefore without any commitment to metaphysical possibility.

However, this strategy will not work. To ideally conceive is to conceive of things being a certain way. If I am not able to describe how things would be, if they were as I conceive them to be, then I am not really conceiving of anything at all; but if I am able to describe how things would be, then I am by definition describing a way things could conceivably be. To ideally conceive is to conceive or suppose that some state of affairs obtains – and *a fortiori* – to conceive of it as possible. The nature of the act of conceiving is essentially modal: it involves a commitment to the possibility of things being a certain way. By its very nature, the act of ideally conceiving entails an epistemic possibility – a way things could conceivably be. It makes no more sense to suppose that we can ideally conceive without conceiving of things being a certain way, than it does to suppose that we can ideally conceive of a way things could not be.

To summarise the argument of this section, the standard formulation of modal dualism is that ideal conceivability entails ideal epistemic possibility, but not metaphysical possibility. But this is straightforwardly incompatible with the uncontroversial definition of metaphysical possibility as unrestricted possibility, or possibility *simpliciter*. Therefore, if there is a coherent formulation of modal dualism, it cannot allow that ideal conceivability constitutes epistemic possibility. It must allow space for ideally conceivable scenarios that do not constitute any form of possibility at all. The only even *prima facie* conceivable way in which this might occur would be if there are ideally conceivable ways things could not be – that is, impossible worlds.

Now, there are two quite different ways to interpret the claim that there are impossible worlds. The first is that there are mere *prima facie* conceptions, such that there does not exist a corresponding possible world. This is unproblematic, but does not satisfy the requirements of modal dualism, which requires an ideally conceivable impossibility. Modal dualism requires that we can ideally conceive of a world W, such that W is impossible. This is what is deeply incoherent. To ideally conceive of a world W is to conceive of it existing as a world; but to conceive of its impossibility is to conceive of its non-existence, since impossible worlds are not really worlds at all. And therefore there is no coherent way to formulate modal dualism – so modal monism is an *a priori* truth.

4.4 - The Problem of Counter-Possible Reasoning

One objection to modal rationalism is that it seems to rule out counter-possible reasoning. That is, it seems to have the consequence that we cannot draw any conclusions about what is entailed by an impossible scenario. Why does modal rationalism seemingly have this consequence? And why would it be a problem? I will answer these two questions, before explaining why, in fact, modal rationalism does not rule out counter-possible reasoning, and does not really pose any such problems.

The problem seems to arise because if ideal conceivability entails metaphysical possibility, then metaphysical impossibility entails the absence of ideal conceivability. Things that are not metaphysically possible are not ideally conceivable. But this appears to mean that necessarily false propositions are not real propositions at all – that they are incoherent or meaningless. And this would be a problem for mathematics, for a start. For example, the proposition that the square root of 2 is rational is a necessarily false proposition, but it is not meaningless. There is no possible world in which the square root of 2 is rational, but we can still understand what is meant by the idea. Indeed, it is precisely because we understand the idea and what it would entail that we are able to do mathematics and prove that root 2 is irrational. It is precisely because we can employ the impossible proposition in our deductive reasoning that we are able to prove that it is impossible. And it is not just mathematics that appears vulnerable to this problem; it seems to apply to *a priori* reasoning generally. Indeed, in this very thesis I have employed counter-possible reasoning to reach my conclusions. I have argued, for example, that *if* modal dualism were true, *then* there would have to be impossible worlds. But how can such reasoning be legitimate if, by my own lights, the very notions of modal dualism and impossible worlds are not really conceivable?

What is the solution to this problem? The answer lies in the distinction between ideal conceivability and mere *prima facie* conceivability, where mere *prima facie* conceivability is the illusion or appearance of conceivability to a limited intellect. In Chapter 3, I suggested that we could also draw a related distinction between ideal propositions and mere *prima facie* propositions: an ideal proposition is one that proposes an ideally conceivable scenario, whereas a merely *prima facie* proposition is one that merely appears, to a limited intellect, to propose a conceivable state of affairs. Now, if modal monism is true, then metaphysical impossibility entails ideal inconceivability. But metaphysical impossibility is perfectly compatible with mere *prima facie* conceivability. Indeed, in order to make sense of the idea that some scenario, S, is metaphysically impossible, we must first have a *prima facie* conception of S.

But what is mere *prima facie* conceivability? I have characterised mere *prima facie* conceivability as being incoherent. And this is right, in the sense that mere *prima facie* conceivability does not present us with a coherent way of conceiving of things as being. There is no coherent way to conceive of the square root of two being rational, for example. But this does not mean that mere *prima facie* conceivability is vacuous or devoid of meaning. On the contrary, mere *prima facie* conceptions routinely contain real and meaningful cognitive contents. The reason they are only *prima facie* conceivable is that they do not combine these genuine contents in a coherent way – so there is, so to speak, no sum of the parts. Thus the *prima facie* proposition that the square root of 2 is rational contains meaningful cognitive contents – namely the concept of rationality, the number 2, and so forth – but it does not in fact combine these concepts to yield a way things could conceivably be, to an ideal conceiver. But the presence of real cognitive contents within the *prima facie* proposition means that we can employ it in deductive reasoning: we can legitimately ask what would follow if the square root of 2 were rational – and thereby prove that it is not.

In summary, modal monism does mean that metaphysically impossible things are not ideally conceivable. It means that necessarily false propositions cannot conceivably be true, which means that they are not ideal propositions. In other words, modal monism entails that necessarily false propositions are merely *prima facie* conceivable. However, this does not mean that counter-possible reasoning is illegitimate, because mere *prima facie* conceivability is not the same thing as meaninglessness. On the contrary, mere *prima facie* propositions – for example, that the square root of two is rational, or that modal dualism is true – routinely contain real semantic content. It is this content that allows us to use such *prima facie* propositions in counter-possible reasoning, and thus show that they are necessarily false.

4.5 - What About the Contingent *A Priori*?

If modal monism is true, then there cannot be any strong *a posteriori* necessities. I have argued that it is indeed true, and that the *prima facie* notion of strong *a posteriori* necessity is, in fact, ultimately incoherent. But modal monism also rules out another apparent class of truth, namely the contingent *a priori*. Contingent *a priori* truths – if any exist – are knowable on the basis of pure reason alone, without recourse to any empirical premise, but are not true in all possible worlds. Why does modal monism rule this out? Because modal monism is the thesis that ideal conceivability and metaphysical possibility are ultimately one and the same thing. Therefore, just as there cannot be an ideally conceivable scenario that is metaphysically impossible, so there cannot be any metaphysically possible world that is not ideally conceivable.

The idea of the contingent *a priori* has received less attention than the necessary *a posteriori*. This is because the inference from ideal conceivability to metaphysical possibility, which is challenged by the apparent existence of necessary *a posteriori* truths, is an essential step in various arguments about the metaphysics of mind, notably the zombie argument against physicalism. The inference in the other direction – from metaphysical possibility to ideal conceivability – plays no such role. Hence the contingent *a priori* seems less metaphysically important; it seems more of a curiosity. It is for this reason that I have focussed on the case of the supposed *a posteriori* necessities in this thesis. But, for the sake of a more complete understanding of modal monism, it is worth looking briefly at the case of the alleged contingent *a priori*.

We should first bear in mind that there is more than one notion of the *a priori*. There are at least two possible definitions. The first definition is broader: a proposition is *a priori* just if it is true for all possible experience. On this definition, our knowledge of the *a priori* does not depend upon any particular experiences, and cannot be refuted by any possible experience; this is the Kantian notion of the *a priori*. It includes those propositions (if there are any such) that are true for all possible experiences, but which are nevertheless metaphysically contingent. So, for example, Kant argued in the *Critique of Pure Reason* that it is knowable *a priori* that there are objects in space and time – but this is certainly a metaphysically contingent proposition (which roughly equates to *synthetic* in Kant's terminology). The existence of such truths is perfectly consistent with modal monism. So there may well be contingent *a priori* truths in this broader, more permissive sense.

But in this thesis I have been using a narrower definition of the *a priori*: a proposition is *a priori* just if its negation is not ideally conceivable. And the Kantian synthetic *a priori* propositions are not *a priori* in this narrower sense. While it may be the case that the existence of objects in space and time is a necessary condition of the possibility of experience, and is thus true for all possible experience, it is certainly not the case that their non-existence is ideally inconceivable. In order for a proposition, *p*, to count as contingent *a priori* in this narrow sense, it must be the case that not-p is both metaphysically possible and ideally inconceivable.

A supposed example is given by Kripke (1980): one metre was originally defined as the length of a particular stick, S, at a particular time, t. Hence it is *a priori* – it can be known without needing to do any measurement, and independently of any particular experience – that the length of S at time t is one metre. But, although it is *a priori* that the length of S at time t is one metre, this is surely contingent. Had S been heated or cooled at time t, it would have been a different length – it would not have been one metre. Hence the truth in question is *a priori*, but contingent.

What is wrong with this? The obvious response is that, just like the Kripkean examples of supposed *a posteriori* necessities, it trades on an ambiguity. So there is a sense in which it is *a priori* that the length of S at time t is one metre, and there is a sense in which it is contingent that S is one metre at t – but there is no sense in which it is both.

It is *a priori* that, whatever the length of S at time t, this shall be known as *one metre*. This is *a priori* because it is true by definition: in fact, it is just a definition of the referring expression 'one metre'. But it involves interpreting the expression 'one metre' in terms of a non-rigid, primary intension: *whatever* the length of S at time t shall be defined as one metre. Hence for any world containing S, if we consider that world as actual (i.e. we interpret the expression 'one metre' according to its primary intension), then it is true by definition that S has length at time t equal to one metre. This is true whatever length S happens to have in the world considered as actual. So therefore it is also necessarily true: there is no possible world in which the length of S at t is not picked out by the expression 'one metre'.

And there is an equally clear sense in which it is contingent that the length of S at time t is on metre. Of course S could have been a different length – in which case, its length would not have been one metre. But this sense requires us to interpret the expression 'one metre' rigidly, in terms of its secondary intension – that is, as referring in all possible worlds to the value that it happens to take in the actual world. In this sense, other possible worlds are considered as counterfactual, and evaluated accordingly. But if the expression 'one metre' is understood this way, then it is *a posteriori* that S has a length at time t equal to one metre. Now, it will still be *a priori* that we should call that length 'one metre', but that is not the point. The point is that the very fact that is contingent – that S at time t has *that particular length* – can only be known *a posteriori*.

So the thing that is contingent and the thing that we can know *a priori* are different. It is contingent that the metre rule has the particular length that it actually has; it is *a priori* that, whatever length it has, that length shall constitute one metre. Thus the Kripkean case of the contingent *a priori* is just as vulnerable to the deflationary two-dimensional analysis as the Kripkean examples of the necessary *a posteriori*.[24]

This raises the question of whether there are cases of the strongly contingent *a priori.* This question has received much less philosophical attention than the question of strong *a posteriori* necessities, for the reasons I have already explained. What I will not do here is go through various candidate cases, as I did for several examples of supposedly strong *a posteriori* necessity. The important point is that the strongly contingent *a priori* leads to similar conceptual problems, and it does so even more directly and obviously.

The problem with strong *a posteriori* necessities is that they require us to posit (in the form of their negations) an ideally conceivable scenario that is not metaphysically possible. This in turn requires us to posit the existence of impossible worlds, which – in the sense required – is itself incoherent. Conversely, the negation of a contingent *a priori* statement would, by definition, be ideally inconceivable, but still metaphysically possible. In other words, it would represent a way things could possibly be, even though they could not conceivably be that way. But this is fundamentally misconceived. As we have seen, to say that a scenario is ideally inconceivable is not to say that there exists a scenario, S, which has the property of inconceivability; it is to say that there does not exist a real scenario corresponding to a mere *prima facie* conception. Just as ideal conceivability constitutes metaphysical possibility, so the absence of ideal conceivability means the absence of anything that would even be a candidate for possibility.

[24] We get the same story if we consider other sources of apparently contingent *a priori* truths. Indexical facts are a case in point – the fact that I am now here can seem like a contingent *a priori* truth, but the deflationary analysis is the same. It is a contingent fact that I am in London at the time of writing, but that is certainly not *a priori*. What is *a priori* is that, wherever I am at the present time shall constitute *here* from my perspective. But this is necessary: there are no possible circumstances in which it is not true.

Strong *a posteriori* necessity posits the existence of a world that has the property of ideal conceivability, but not the property of being metaphysically possible. This is misconceived, because if there is no metaphysical possibility, then there is no world. Similarly, the contingent *a priori* posits the existence of a world with the property of being metaphysically possible, but not the property of being ideally conceivable. This is similarly misconceived, because without ideal conceivability, there is no world at all. So the very notion of a metaphysically possible world that is not ideally conceivable is incoherent.

Conclusion

In this chapter I have contrasted two opposing ways of understanding the fundamental nature of possibility and necessity. According to modal monism, metaphysical possibility and ideal epistemic possibility are ultimately the same thing: both are constituted by ideal conceivability. Similarly, the metaphysical necessity of a proposition is constituted by the ideal inconceivability of its negation. Modal dualism, on the other hand, asserts that there are at least two fundamental types of possibility: an epistemic notion, constituted by ideal conceivability; and a metaphysical notion, which is independent of considerations of conceivability and the *a priori*.

Modal dualism thus presents the possibility that there are strong *a posteriori* necessities – that is, that there are propositions whose negations are ideally conceivable, but not metaphysically possible. Modal monism rules out the possibility of strong necessities, since every ideally conceivable scenario must constitute a metaphysically possible world.

The debate between these two fundamentally different views of modality is often presented as a contest between two competing, but equally coherent theories. It often seems to be a question of which theory is best able to explain the modal data, whilst drawing on the fewest brute, unexplained modal facts. But I have argued that this is a misconception of the real, underlying issue. The real issue is that modal dualism itself is deeply incoherent. Specifically, the notion of an ideally conceivable scenario that is not metaphysically possible entails the existence of impossible worlds, which is an incoherent notion. Moreover, even allowing the possibility of *a posteriori* metaphysical necessities would render us unable to distinguish between the metaphysically necessary and the merely contingent. So we should conclude that *a posteriori* necessities are impossible, and that ideal conceivability does entail metaphysical possibility.

Chapter 5: Zombies Redux

Introduction

Having set out the case for modal rationalism over the last three chapters, I will now return to the central argument against physicalism: the zombie argument. The simplest form of the argument, as set out in Chapter 1, is as follows:

i) If zombies are metaphysically possible, physicalism is false.

ii) Zombies are ideally conceivable.

iii) The ideal conceivability of a scenario entails that there is a corresponding metaphysical possibility.

Therefore:

iv) Zombies are metaphysically possible.

Therefore:

v) Physicalism is false.

As we have seen, the main objection to this argument is to reject premise (iii) – that is, to deny that conceivability entails possibility. I have shown that this objection is misplaced, and argued that ideal conceivability does indeed entail metaphysical possibility. But that is not the only possible objection; in this chapter, I will consider several others. What other possible objections might supporters of physicalism make? As I argued in Chapter 1, premise (i) follows from the definition of physicalism in terms of metaphysical supervenience, so I will take it that this premise is not in question. However, this still leaves several options. The first is to deny the validity of the argument. The second is to deny that zombies are really conceivable – to claim, in fact, that they are merely *prima facie* conceivable, but are not ideally so. So in Section 5.1, I will consider an objection to the validity of the argument, and show that this objection does not succeed – the argument is valid; then, in Sections 5.2 and 5.3, I will consider objections to premise (ii), the conceivability of zombies. Section 5.4 deals with a specific problem raised by the doctrine of *dispositional essentialism*. In Section 5.5, I will consider a novel objection to the argument, concerning how we should evaluate zombie judgements. I will conclude that all these objections can be defeated, and therefore the argument succeeds.

5.1 - The Validity of the Zombie Argument

Even if, as I have argued, premises (i), (ii), and (iii) of the zombie argument are true, a further concern is that the argument as it is set out above is not valid. One possible objection runs as follows: premise (i) is ambiguous, since it is only the 2-possibility of zombies that is incompatible with physicalism; but premise (ii) only establishes the 1-conceivability of zombies; therefore, even if premise (iii) is conceded, the most it would establish is their 1-possibility; therefore the argument is not valid, and does not show that physicalism is false. It is worth noting that this objection is not merely a re-statement of the previous objection that conceivability does not entail possibility. The present objection is that there is a difference between the thing of which I can conceive and the thing whose possibility would be incompatible with physicalism: the thing of which I can conceive is the 1-possibility of zombies, but the thing that is incompatible with physicalism is their 2-possibility. And modal rationalism does not imply there is a straightforward, general entailment from 1-conceivability to 2-possibility. The entailment from conceivability to possibility only applies when the thing that is conceivable is the same as the thing that is possible.

I will show that it is relatively straightforward to reformulate the zombie argument in a way that avoids this objection. The objection does raise a valid point, which is that it is the 2-possibility of zombies – and not merely their 1-possibility – that is required to refute physicalism. However, the objection fails because, although there is no general entailment from 1-conceivability to 2-possibility, the entailment does apply in the specific case of zombies.

Why does the zombie argument need to show that zombies are 2-possible? The most obvious answer is that physicalism will be false just if zombies are metaphysically possible, and it is natural to identify metaphysical possibility with 2-possibility. Hence Chalmers writes:

> […] materialism requires not the 1-impossibility of P&~Q but the 2-impossibility of P&~Q. That is, materialism requires that it *could not have been the case* that P is true without Q also being true. This is a subjunctive claim about ordinary metaphysical possibility **and so invokes 2-impossibility rather than 1-impossibility**. [2010, p149. Author's italics; my emphasis.]

Here Chalmers identifies metaphysical possibility with 2-possibility – which would indeed mean that, for the zombie argument to work, it is necessary to show that zombies are 2-possible as well as 1-possible.

However, I do not think this is quite right. For the reasons set out in Chapter 4, I think it is misleading to identify metaphysical possibility with 2-possibility. Nonetheless, I do think that, the zombie argument still requires zombies to be 2-possible, albeit for different reasons. The reason stems from the specific claim made by physicalism: it claims that the actual physical facts are metaphysically sufficient for consciousness. It does not claim that any possible physics would be metaphysically sufficient for consciousness; merely that the actual physics happens to be so. Accordingly, zombies are supposed to be physical doppelgangers of ourselves, whatever the truth of physics turns out to be. Whatever our physical properties may ultimately turn out to be, we want zombies to be exactly like *that*. So physicalism is compatible with the 1-possibility of zombies (which might look like us, but have a very different underlying physics), but not their 2-possibility.

This is represented in the two-dimensional matrix below, for the statement 'there are zombies'. W1 is the actual world – just us, no zombies. W2 is an alternative possible world that superficially resembles our own world. It contains conscious beings that are psychologically qualitatively identical to us, and which at face value resemble us physically – but they are different from us at the level of fundamental physics, in a way that we cannot discriminate with our current understanding of physics. WZ1 is the zombie world: the beings there are exactly like us physically (and therefore not exactly like the inhabitants of W2), but there is nothing it is like to be them. WZ2 is the equivalent zombie world for the inhabitants of W2. To us, the inhabitants of WZ2 are not true zombies – they are just inanimate objects that superficially look like us; and to the inhabitants of W2, the creatures in WZ1 are not true zombies.

‘There are zombies’		World considered as **Counterfactual**			
		W1	**W2**	**WZ1**	**WZ2**
World considered as **Actual**	**W1**	F	F	**T**	F
	W2	F	F	F	T
	WZ1	F	F	T	F
	WZ2	F	F	F	T

The bottom line is that, in order for the zombie argument to work, there needs to be a possible world WZ1 that is the zombie equivalent of the actual world, W1. When WZ1 is considered as counterfactual from the perspective of the W1, it yields the value ‘true’ for the statement ‘there are zombies’ – that is, from the perspective of us human inhabitants of W1, zombies are 2-possible.

So the argument against physicalism needs to show that zombies are 2-possible. But does the 2-possibility of zombies follow from their 1-conceivability? The 1-possibility of zombies is supposed to follow from their 1-conceivability – so what is needed (Chalmers argues) is a premise that if they are 1-possible, then they are 2-possible. The result is an amended version of the argument (cf 2010, p149), as follows:

i) If zombies are 2-possible, physicalism if false.

ii) Zombies are 1-conceivable.

iii) If zombies are 1-conceivable, zombies are 1-possible.

iv) If zombies are 1-possible, zombies are 2-possible.

Therefore:

v) Physicalism is false.

The issue is now premise (iv): does the 1-possibility of zombies entail their 2-possibility? Chalmers argues (2010, p149) that the 1-possibility of zombies entails their 2-possibility if both the physical and mental properties involved in the zombie scenario (abbreviated as P and Q, respectively) have primary and secondary intensions that coincide.

Chalmers argues that it is highly plausible that primary and secondary intensions do coincide in the case of phenomenal concepts (i.e. those concepts that refer to conscious states). This is an important point: for the argument to work, it is no good if our phenomenal concepts have secondary intentions that pick out whatever underlying physical structure instantiate consciousness, in the same way that H_2O instantiates the watery properties. These physical structures, after all, will be present in the zombie universe just as much as the non-zombie world. So if these structures are the real referents of our phenomenal concepts, then any physical duplicate of the actual world will automatically contain consciousness. But we need not worry about this point too much. It is clear that our phenomenal concepts do not work like, for example, our 'water' concept. Referring to Kripke, Chalmers notes that 'there does not seem to be the same strong dissociation between appearance and reality in the case of consciousness as in the cases of water and heat' (p149). When we evaluate a counterfactual world for the presence of consciousness, the important thing is whether or not the relevant subjective feelings are instantiated: if something feels like pain, then it is pain; and if it doesn't feel like pain, then it isn't – regardless of the underlying physical structures.

However, Chalmers then acknowledges (p150) that it is less plausible that the primary and secondary intensions of the physical properties coincide. It may be, he concedes, that the structural roles of mass, charge, and so forth, are played by as yet unknown intrinsic properties of matter. Perhaps, if we knew what *these* were like, we would see that there is no 2-conceivable world in which *these* intrinsic properties were instantiated and yet consciousness was not. So, although zombies may be 1-possible, perhaps they are not 2-possible. But, Chalmers then argues, this scenario would represent a form of panpsychism or panprotopsychism: at least some of the intrinsic fundamental properties would need to be mental or proto-mental to entail *a priori* the emergence of consciousness; if there were no intrinsic mental or proto mental fundamental properties then it would be easy to construct a parallel conceivability argument involving *intrinsic zombies*. But if we disregard this possibility – and, for the purposes of this thesis, I have assumed that the fundamental properties of the physical world are non-mental, which effectively rules this possibility out – then the argument will go through and zombies will be 2-possible.

Whilst I think this argument ultimately works, I also think it is unnecessary – for there is an easier way to get the desired conclusion. Instead of arguing from the 1-possibility of zombies to their 2-possibility, we could argue quite straightforwardly from their 1-conceivability to their 2-conceivability, which would then entail their 2-possibility. This would give us the following version of the argument:

i) If zombies are 2-possible, physicalism is false.

ii) Zombies are 1-conceivable.

iii) If Zombies are 1-conceivable, zombies are 2-conceivable.

iv) If zombies are 2-conceivable, zombies are 2-possible.

Therefore:

v) Physicalism is false.

Now the crucial premise is (iii). But why does the 1-conceivability of zombies entail their 2-conceivability? Simply because, once we have made the assumption that the underlying physics of the actual world is non-mental or proto-mental, thus ruling out any form of panpsychism or panprotopsychism, it does not matter what the physics turns out to be. Whatever the physics of actual humans is like, we can just stipulate that zombies are exactly like *that*, but without consciousness. We can be certain that there are no lurking contradictions between the actual physics and the zombie scenario, because we have already ruled out panpsychism and its variants for the sake of argument. There is no requirement for us to have a positive 2-conception of zombies: their ideal negative 1-conceivability entails that they are 2-conceivable, even if, not knowing the physics of our own world, we are not in a position to formulate the 2-conception in detail.

5.2 - Are Zombies Really Conceivable?

Another way to defend physicalism from the zombie argument is to deny premise (ii) – that is, to deny that zombies are really conceivable. This will result in what Chalmers categorises as *a priori,* or Type-A physicalism.

The zombie scenario *seems* conceivable, at least at face value. But perhaps, when we examine the idea in detail, we will find that it is incoherent. As I have argued, there is a distinction between *prima facie* conceivability – that is, what seems, on first appearances, to be conceivable to me, with my limited intellectual powers – and *ideal* conceivability – that is, what would be conceivable to an infinitely powerful idealised rational thinker. Thus a further objection to the zombie argument is that zombies are only *prima facie*, and not ideally conceivable. There is no ideally conceivable minimal physical duplicate of the actual world that is not a duplicate with respect to consciousness. In other words, *a priori* physicalism asserts that there is an *a priori* entailment from the physical facts of the actual world to the facts about consciousness.

What does it mean to assert that there is an *a priori* entailment from the physical facts to the conscious facts? According to Chalmers (2010, p111) it means that there is no epistemic gap between the two sets of facts, at least from the point of view of an ideal rational conceiver. In other words, if an ideal rational conceiver were in a position to know the physical facts, P, then they would be able to deduce the conscious facts, Q, with no further knowledge. An ideal conceiver would not require any additional knowledge precisely because they would know it to be inconceivable that P and not-Q; thus, knowing P, they would immediately know that Q. According to Chalmers, then, the claim that zombies are conceivable is just the claim that, from the point of view of an ideal conceiver, there would be an epistemic gap between the physical facts and the conscious facts; conversely, the claim that zombies are not really conceivable is just the claim that there is no such gap.

If *a priori* physicalism is true, then it does not merely rule out the conceivability of a zombie world: it would also rule out the conceivability of any world that is a physical duplicate of our own, but which differs from ours with respect to its conscious properties in any way whatsoever. *A priori* physicalism does not merely rule out *total* zombies (which there is nothing it is like to be): it also rules out *partial* zombies, which are physical duplicates that lack *some* of our conscious states; and it rules out *inverts* (Chalmers 2010, p154), which are physical duplicates whose conscious properties are the same as ours, but are rearranged with respect to the underlying physical properties (so an invert might have the subjective experience of red where its physical doppelganger has the experience of blue, and blue in place of red). If *a priori* physicalism is true, then *any* conceivable universe that is a minimal physical duplicate of the actual world must *a priori* be an *exact* duplicate with respect to consciousness.

Given that zombies do seem to be *prima facie* conceivable – there is no obvious contradiction in the scenario that I described above – then the burden of proof falls on the *a priori* physicalist to explain why the appearance of conceivability is deceptive. In this section, I will argue that the appearance is *not* deceptive; and that the arguments in favour of *a priori* physicalism are themselves ill-conceived. Before I do so, I will outline several issues that I believe are not relevant to the central argument – and which represent, in my view, inconsequential or weak arguments for *a priori* physicalism.

The first relates to another distinction, which I briefly outlined in Chapter 2, that Chalmers draws between two types of conceivability – in this case, negative and positive conceivability (2010, p144). Positive conceivability means that it is possible to imagine a scenario in full, with no gaps, so to speak, in our mental picture of what it would be like. Negative conceivability, on the other hand, is a weaker notion: it means that we can imagine the defining features, the outline of a scenario, and we can see that there is no inherent contradiction in it – but we may not be able to imagine in full what the scenario would be like. For an ideal rational conceiver, this distinction would be irrelevant – a scenario would either be conceivable, or it would not. The distinction only arises for beings of finite power, who may not be able to imagine a scenario in full detail, but can nonetheless form an outline conception of it. For a limited rational being, if a scenario is positively conceivable, then it must also be negatively conceivable; but negative conceivability does not entail positive conceivability.

One misconceived objection to zombies is that they are merely negatively, and not positively conceivable (cf 2010, p157). But I think this objection misses the point, for two reasons. First, it seems plain enough that zombies are *prima facie* negatively conceivable. If this is right, then the proper question is whether or not they are ideally conceivable. If they are ideally negatively conceivable, then the fact that we cannot positively conceive of them is no objection; and if they are not ideally negatively conceivable, then *a fortiori* no rational being could form a positive concept of them. Either way, positive conceivability does not add anything. Second, as Chalmers argues, the positive conceivability of zombies is irrelevant because it is precisely the *absence* of consciousness that we are invited to imagine: there is just nothing it is like to be a zombie, and there is, so to speak, no more detail to be filled out. As Chalmers puts it:

> There is no more problem with clearly and distinctly imagining a situation in which there is no consciousness than in imagining a world in which there are no angels or in imagining a world with one particle and nothing else. The argument here appears to require that absences are never positively conceivable or at least that to positively conceive an absence always requires conceiving something else in its place. But these cases suggest that such a claim is clearly false. [2010, p 157]

A similar and equally misplaced objection is that we could never be sure that zombies are ideally conceivable, because we are not ourselves ideal conceivers. But this does not follow. Again, as Chalmers argues (2010, p155), even though we are not ourselves ideal conceivers, we know that the rational square root of 2 is not ideally conceivable, and that (for example) a flying horse or creature with the body of a lion and the head of an eagle are both ideally conceivable. In order to know whether or not something is ideally conceivable, we do not have to be ideal rational conceivers ourselves – if that were the case, we would never be able to know *anything a priori* – we just need sufficient imagination and reasoning power to be certain whether or not, in a particular case, an idea is self-contradictory. There may indeed be some cases that are marginal, and which are beyond the abilities of mere humans to settle one way or the other. Some mathematical ideas may fall into this category; and it may be that zombies will ultimately do so – but the mere fact that we are not ideal thinkers does not make this inevitable.

In summary, the issue is whether or not there is an *a priori* entailment from the physical facts in the actual world to the conscious facts of the actual world. But is there any such entailment? What would such an entailment look like?

One possibility, which I will not consider in this thesis, is the idea that the epistemic gap between the physical facts and consciousness could be closed by expanding our conception of the physical. The idea is that zombies only appear to be conceivable because we have an inadequate understanding of the physical facts in the actual universe; if we were able to truly understand the physical world, then we would see that consciousness follows *a priori*, that it is not something additional to the physical facts. If some form of panpsychism or panprotopsychism is true, then something like this picture will be right. But for the purpose of this thesis I have already discounted panpsychism and panprotopsychism. As I set out in Chapter 1, I will make the assumption that the physical world – whatever it may ultimately be like – is non-mental and does not contain any proto-mental properties at its most fundamental level. So, whilst it may indeed be the case that we could close the epistemic gap by radically expanding our concept of the physical, I will not consider this possibility. The issue at present is whether or not it is possible to close the epistemic gap *without* radically expanding our concept of the physical.

The view that there is an epistemic gap between the physical facts and the facts about consciousness is widely accepted, in large part due to Jackson's (1982) *knowledge argument*. This thought-experiment imagines a scientist, Mary, who investigates the world from a black-and-white room, exposed to no other colours. She acquires a complete understanding of the physics and neurophysiology of human colour perception, to the extent that she has a perfect knowledge of the physical facts involved in human perception of the colour red. She then leaves the room and sees a red object. In doing so, she has learnt something that she could not know, even with perfect physical knowledge, and even with idealised rational powers, on the basis of the physical facts alone – namely what it is like to see red. This means that the facts about what it is like to see red – that is, the facts about consciousness – are not *a priori* entailed by any of the facts that Mary had access to in the black and white room. There is an epistemic gap between the physical facts on one hand and the phenomenal facts on the other. In Jackson's terms, the complete physical information regarding colour-perception that Mary acquired whilst in the black and white room did not enable her to know what it is like to see red; on seeing a red object, she acquires new information – information which is therefore not physical. Jackson (1982) takes this to be a refutation of physicalism.[25]

[25] Jackson has subsequently changed his position. In 1998c, he rejects his previous view that phenomenal consciousness is non-physical. This is because anti-physicalism supposedly leads to epiphenomenalism, which is an unacceptable consequence. Mary's knowledge of what it is like to see red is fully explicable in physical terms (because, according to Jackson, interactionist dualism is false); and it is not plausible that the content of Mary's knowledge – the thing that she comes to know when she see red – outruns the causal explanation of how she acquires this knowledge. So the thing that she comes know must in fact be a physical fact. How then, in Jackson's revised view, to explain why it *seems* that Mary learns something new that outruns what can be deduced from the physical facts? Jackson's answer is that sensory experience is an intrinsically representational way of acquiring,

One possible response to the knowledge argument is that, even if it succeeds in establishing an epistemic gap between the physical facts and the facts about consciousness, this does not entail that there is a metaphysical gap. For example, Horgan (1984) argues that the knowledge argument trades on an ambiguity in the notion of *physical information*. There is a distinction, he argues, between *explicitly physical information* and *ontologically physical information*: the former is information that is expressed in physical terms; the latter is information about physical things. Although, on seeing a red object for the first time, the new information acquired by Mary – about what it is like to see red – is not explicitly physical, this does not entail that it is not ontologically physical. Mary may have acquired new knowledge that cannot be described in physical terms – but it is knowledge of an old, physical fact nonetheless. Now of course the central question of this thesis is whether an (ideal) epistemic gap entails a metaphysical one, so I will not dwell on the matter here.

The more pressing question at present is whether Jackson's argument actually establishes an epistemic gap between physical facts and facts about consciousness. Another possible objection to the knowledge argument is that Mary does not really gain new knowledge when she is released from the black-and-white room. Rather she has the same knowledge as before – knowledge of physical facts about the reflection and absorption of light by external objects, her own neurological states, and so forth – but knows all this in a new way, under a new phenomenal mode of presentation. When in the black and white room, she knew the physical facts by description; upon her release she acquired the same knowledge by direct acquaintance in her colour-perception.

The problem with this objection is that, even on its own terms, it does not refute the claim that Mary gains substantive new knowledge on her release. Even if it is granted that she gains old knowledge (of physical facts) in a new way (by perception), the very fact that she can know

very speedily, information about the relational and functional properties of the world, including our own internal states.

I am more sympathetic to Jackson's (1982) argument than his subsequent recantation. The epiphenomenalism issue represents a real problem for anti-physicalist views of consciousness. In Chapter 1, I briefly outlined the causal argument for physicalism; I will not offer a solution here. However, I am not persuaded by the later Jackson's explanation of the intuition that there is an epistemic gap – it leads in my view to an *a priori* version of physicalism that is little more than the denial of subjective consciousness.

the old facts in a new way is itself a substantive piece of knowledge, which was not accessible to her in the black and white room. The fact that she can know physical facts in the new way is not contained within or entailed by the descriptive knowledge that she had in the black and white room (cf Chalmers 2010, pp 195-196).

Therefore Jackson's (1982) knowledge argument does establish at least a prima facie epistemic gap between the physical facts and the phenomenal facts. But can this apparent epistemic gap be closed? One proposal, considered by Chalmers (2010, pp113-114), is that some form of functionalism can bridge the epistemic gap. The idea requires two things to be true: first, that certain types of mental states – in particular, mental representations – can be given a functional analysis, such that any physical system which instantiates the relevant functions must *a priori* realise the mental states in question; second, that these functionally constituted mental states are *a priori* sufficient for consciousness. If this is correct, then there will be an *a priori* entailment from the physical facts to the relevant functional roles; and an *a priori* entailment from the functional roles to the representational mental states; and an *a priori* entailment from the representational mental states to consciousness; so the epistemic gap will be closed.

Now, it does seem plausible that something along these lines is true of macroscopic physical phenomena, such as biological life. In this case, the microphysical facts are sufficient to determine *a priori* the functional facts about reproduction, metabolism and so forth – and the existence of biological life itself is nothing over and above the instantiation of these functions by appropriate physical structures. Functionalism about biological life therefore does not require that life can be analytically defined in microphysical terms – merely that it can be functionally defined, and that microphysical structures can play the relevant functional roles.

The analogy between macroscopic phenomena such as biological life and functionalism with respect to consciousness is an important one. If *a priori* physicalism is to be true, this analogy needs to hold. It is precisely because functionalism is true of biological life that we cannot ideally conceive of a world that is a physical duplicate of our own, but which differs with respect to the facts about life. The whole point of type-A physicalism is just that consciousness is not something special and different, but is a physical phenomenon just like any other, and can be understood along the same lines.

But the analogy between biological life and consciousness breaks down on closer inspection. The problem for the analogy is that the concept of representation is ambiguous: on one hand it can refer to a conscious representation, that there is something it is like to have; on the other, it can refer to something like a functional information-processing state of a non-conscious machine. As Chalmers puts it:

> On examination though, this argument appeals to an ambiguity in the notion of representation. There is a notion of *functional representation*, on which *p* is represented roughly when a system responds to *p* and/or produces behaviour appropriate for *p*. In this sense, explaining functioning may explain representation, but explaining representation does not explain consciousness. There is also a notion of *phenomenal representation*, on which *p* is represented roughly when a system has a conscious experience as if *p*. In this sense, explaining representation may explain consciousness, but explaining functioning does not explain representation. [2010, p 114; author's italics.]

Why is this a problem for *a priori* physicalism? Because the physical facts *a priori* entail representation only in the functional sense; and consciousness is *a priori* entailed only by representation in the phenomenal sense. Compare this with the equivalent type-A physicalist claim about biological life: the physical facts *a priori* entail the facts about biological functions (reproduction, metabolism etc); and the same facts about biological functions *a priori* entail the existence of life. In the biological case, there is no ambiguity: biological functioning entails the existence of life, unambiguously. But functional representation does not entail the existence of consciousness – only phenomenal representation does.

But what motivates the distinction between mere functional representation and phenomenal representation? The answer is the subjective character of consciousness – the *what it is like* - that is by definition present in phenomenal representation and absent in mere functional representation. Having a functional analysis of representation does not help: we are still in the same position relative to *conscious* representation as Mary, while still in her black and white room, was to the experience of seeing red. Even if we suppose that Mary, as well as being expert on colour-perception, also has complete knowledge of the neurological implementation of human cognitive functions – so she has a perfect understanding of how our brains track external objects, how they adjust system outputs (including behaviour) accordingly, how our tracking systems can monitor their own states, and so forth – there would still be no *a priori* entailment from this information to facts about conscious representation, because the functional description of these states does not entail that there is something it is like to have them. Text-book knowledge of information-processing in the human brain will not tell her what it is like to have conscious thoughts (though of course, she knows this from her own subjective consciousness, even while still in the black and white room).

So, in conclusion, the idea that zombies are not really conceivable rests upon an analogy between consciousness and macroscopic physical phenomena such as life, which is only plausible if we ignore the subjective character of consciousness – that is, if we ignore consciousness entirely.

5.3 - Are Zombies Really Really Conceivable?

A novel objection to the conceivability of zombies is presented in Giberman (2015). The target of the argument is specifically the 1-conceivability of zombies; for the sake of developing the argument, Giberman assumes that ideal conceivability entails metaphysical possibility. If the argument succeeds, then it would refute the first premise of the overall two-dimensional argument against physicalism, and that argument would be defeated. Moreover, if zombies are not 1-conceivable, then it would follow that they are not 2-conceivable either. It wouldn't matter what actual value is assigned to *P*, the physics of the actual world: *whatever* the physics of the actual world, it could not conceivably be such as to permit zombies.[26]

Before setting out the argument, Giberman outlines two preliminary considerations. The first (pp 124-125) is the concept of a Mereological Threshold for Consciousness (MTC). The idea here is that conscious entities, such as humans, are typically complex entities composed of many fundamental parts; these fundamental parts are termed *quarks* for the purpose of the argument. Between an individual quark and a complete human being are many different complexes of quarks, each of which is a proper part of the complete human being. In the actual world, some of these complex entities would be capable of being conscious if they were detached from the complete human being, whilst others would not. Thus, in the actual world,

[26] There is a more general point here: that 1-inconceivability entails 2-inconceivability. If it is inconceivable that watery stuff in general has some property, *p*, then it is inconceivable that *this particular* watery stuff (whatever it may turn out to be) has property *p*.

an individual quark detached from my body would not be capable of being conscious; nor would my left hand. But the mereological sums consisting of me-minus-my-undetached-left-hand, or me-minus-one-random-undetached-quark would be capable of consciousness, if they were detached. Thus there exists a spectrum of mereological sums, with a complete human being at one pole, and each individual quark at the other, and all possible combinations in between. Somewhere along this spectrum will be a threshold – the MTC – such that every mereological sum on one side of the threshold would incapable of consciousness if it were to be detached, and every mereologoical sum on the other side would be capable of consciousness if detached. The addition or subtraction of a single quark (correctly placed) would cause an entity to cross the MTC from one side to the other. An obvious objection to this idea is that there may not be an exact threshold between consciousness-capable entities and consciousness-incapable ones. It seems unlikely that the addition of a single quark in the right place would suddenly cause consciousness to appear, like a light being switched on; it seems more likely that the boundary is fuzzy. But, as Giberman argues (p 137), it does not really matter whether the boundary is fuzzy or exact. The point is that, in the actual world, individual quarks are not capable of consciousness, complete human beings are, and, somewhere in between, a threshold of physical complexity is crossed, and there will be a transition from entities that are not capable of consciousness to those that are.

The second preliminary consideration is that, according to Giberman, the 1-possibility of panpsychism is incompatible with the 1-conceivability of zombies. Therefore if panpsychism is 1-possible[27] then zombies are not 1-conceivable. The argument for this principle worth quoting in full:

> Physical panpsychism is the thesis that phenomenal consciousness is an intrinsic categorical property of mereologically basic particulars, which plays a constitutive, underwriting role in (i) the fundamental properties of ‘final’ physics in the actual world and (ii) the exemplification of consciousness by more complex structures. To see that the primary conceivability of zombies is inconsistent with the primary possibility of physical panpsychism, notice that physical panpsychism has actuality built in: it is a thesis about actual final physics. So even the primary possibility of physical panpsychism would entail that actual physics presupposes consciousness. Consequently, one cannot coherently conceive of a state of affairs that is physically indiscernible from the actual world – as required by the primary conceivability

[27] Or 1-conceivable. Although Giberman does not do so in his argument, I will use conceivability and possibility interchangeably, since ideal conceivability entails metaphysical possibility and vice versa.

> of zombies – unless either the physical structures in that state of affairs are conscious or physical panspychism is assumed primarily impossible. [p 128]

With these preliminaries in mind, we can now set out the main argument (pp 129-130), which can be summarised as follows:

> i) Any actually conscious physical structure, X, has a MTC. [From preliminary considerations]
>
> ii) Either it is ideally 1-conceivable or it is not 1-conceivable that the MTC of X could have been different. [Tautology].
>
> iii) If it is not ideally 1-conceivable that the MTC of X could have been different, then zombie versions of X are not 1-conceivable. [Follows from definitions of MTC and zombies].
>
> iv) If it is ideally 1-conceivable that the MTC of X could have been different, then it is ideally 1-conceivable that physical panpsychism is true.
>
> [The argument (p 129) for this premise is that, if it is conceivable that the MTC could have been different than it actually is, then it is conceivable that it could occur anywhere along the spectrum from a single quark at one pole to a complete X at the other. Thus: '…all physical spatiotemporal-cum-mereological structures are equal with respect to being imaginably and coherently consciousness-capable.' But if it conceivable that a single quark is consciousness-capable, then it is conceivable that physical panpsychism is true.]
>
> v) If panpsychism is ideally 1-conceivable, then it is 1-possible. [Assumed]
>
> vi) If panpsychism is 1-possible, then zombies are not 1-conceivable. [From preliminary considerations]

Thus, whichever horn of the dilemma (in premise (ii)) is taken, it follows that zombies are not 1-conceivable. Therefore zombies are not 1-conceivable.

What should we make of this argument? My view is that premises (i) to (v) are true, and the argument is valid. Obviously, I do not accept the conclusion. The problem lies with premise (vi): contra Giberman, it is not the case that the 1-conceivability of zombies is inconsistent with the 1-possibility of panpsychism. Clearly, there is no ideally conceivable zombie-world in which panpsychism is true. While panpsychism itself is metaphysically possible and is therefore true at some worlds, a world where panpsychism is true cannot contain zombies (because, excluding blockers, the intrinsic properties of the physics of that world will be metaphysically sufficient for consciousness). So it is not 1-conceivable that: there are zombies and panpsychism is true. But this is not the premise that Giberman needs; he needs the stronger premise that if panpsychism is 1-possible, then zombies are not 1-conceivable. But what is the justification of this claim?

The justification is supposed to come in the preliminary considerations. The key move is the following line in the above-quoted paragraph:

> panpsychism […] is a thesis about actual final physics. *So even the primary possibility of physical panpsychism would entail that actual physics presupposes consciousness.* [My italics]

But this is not right. The primary possibility of panpsychism does not entail that actual physics (understood rigidly) presupposes consciousness, but merely that it is possible that actual physics (whatever it may turn out to be) presupposes consciousness.

Now, it is true that the 2-possibility of panpsychism is incompatible with the 2-conceivability of zombies. Why is this? Because there is no real difference between the claim that panpsychism is 2-possible and the claim that it is true. Actual physics either presupposes consciousness or it does not; if the actual facts are such that they themselves (as opposed to some other possible physics) possibly presuppose consciousness, then they simply presuppose consciousness. But if actual physics presuppose consciousness, then zombies are not 2-conceivable.

But is there any entailment from this fact – that the 2-possibility of panpsychism is incompatible with the 2-conceivability of zombies – to the premise that Giberman needs? Does it follow that the 1-possibility of panpsychism is incompatible with the 1-conceivability of zombies, as per premise (vi)?

In order to deduce premise (vi) we would need the following two additional premises: first, that the 1-possibility of panpsychism entails the 2-possibility of panpsychism; second, that the 1-conceivability of zombies entails the 2-conceivability of zombies. Armed with these additional premises, we could reason as follows: if panpsychism is 1-possible, then it is 2-possible; if panpsychism is 2-possible, then zombies are not 2-conceivable (by the reasoning given above); but if zombies are not 2-conceivable, then they are not 1-conceivable (which follows from the second additional premise); therefore if panpsychism is 1-possible, then zombies are not 1-conceivable. In this case, premise (vi) would be true, and his argument would succeed.

But the additional premises are false. The 1-possibility of panpsychism does not entail its 2-possibility. If we are agnostic with respect to final physics, then it is possible that final physics turns out to be such that panpsychism is true; but this does not entail that final physics is such that it possibly presupposes consciousness. Similarly, there is no entailment from the 1-conceivability of zombies to their 2-conceivability – unless we discount the possibility of panpsychism. But of course, if we discount the possibility of panpsychism, then Giberman's whole argument collapses.

So there is no entailment from the true claim that the 2-possibility of panpsychism is incompatible with the 2-conceivability of zombies, to the required premise (vi). For this reason, Giberman's argument fails, and there is no reason to suppose that zombies are not conceivable.

5.4 - Dispositional Essentialism and the Possibility of Zombies

Another apparent problem for the zombie argument comes from the doctrine of *dispositional essentialism.* The central idea of this doctrine is that fundamental physical properties are constituted by their causal dispositions. It follows that causal dispositions of fundamental properties are essential to them (see e.g. Choi & Fara, 2018). Suppose, for example, that mass is a fundamental physical property, and that the fundamental causal dispositions associated with mass are given by the general theory of relativity. Then if dispositional essentialism is true, it will be essential to mass that its causal dispositions conform to general relativity. So, while there may be possible worlds that contain mass-like properties that do not conform to general relativity, there are no possible worlds containing mass itself but in which general relativity is not true. If it doesn't conform to general relativity, then it doesn't count as mass.

Dispositional essentialism stands in contrast to *categoricalism*, which asserts that the essence of a fundamental property is constituted by its intrinsic nature. It may be that this intrinsic nature entails certain causal roles in certain circumstances, but the important point is that the dispositional roles themselves are not essential. I will not attempt to settle the dispute between dispositional essentialism and categoricalism here. What I will address here is the concern that dispositional essentialism, if true, would undermine the zombie argument.

Why does dispositional essentialism appear to threaten the zombie argument? To see why, suppose that it is a fundamental causal law that, in the right circumstances, the fundamental physical properties of the world give rise to consciousness – that is, we assume that there are psycho-physical connecting laws of nature. In this case, one of the fundamental dispositions of the fundamental physics of the world will be to give rise to consciousness (under the right conditions). But, if dispositional essentialism is true, then this causal disposition will be essential to the fundamental physical properties of the actual world. Therefore, any possible world that is a complete physical duplicate of the actual world must also duplicate the disposition of fundamental physics to cause consciousness. If it doesn't cause consciousness, then it isn't a duplicate of actual physics. And this means that there is no possible world that is a complete and minimal physical duplicate of the actual world, but which does not contain consciousness – there is no zombie world.

So it appears that dispositional essentialism is fatal to the zombie argument. Dispositional essentialism plus the existence of psycho-physical connecting laws of nature seemingly entails that there is no possible zombie world. But is this right? Is it necessary to reject dispositional essentialism about fundamental physical properties, if we want to rescue the zombie argument?

In my view, the problem is more apparent than real. The apparent problem is that dispositional essentialism plus psychophysical connecting laws entails that actual physics necessitates consciousness. But the solution is that this is only a weak *a posteriori* necessity. Dispositional essentialism means that our physical concepts rigidly designate the fundamental causal roles that physical properties play in the actual world. Any alternative possible physics that does not instantiate the same fundamental causal roles will not count as instantiating the same fundamental properties. So, if dispositional essentialism is true and there are psycho-physical connecting laws, then it will be necessary that physics causes consciousness in the same way that it is necessary that water is H_2O. If dispositional essentialism is true and there are psycho-physical connecting laws, then no possible physics will count as *our* physics unless it also causes consciousness. In this case, whilst there will not be a possible world that is a complete physical duplicate of the actual world – including the disposition to cause consciousness – but which lacks consciousness, this is hardly surprising. Any physics that does not give rise to consciousness will simply not count as a duplicate of our physics. But of course, there will still be possible worlds that are duplicates of the actual world with respect to all the other fundamental dispositional roles – the mass role, the charge role, and so forth – but which lack consciousness. And it will still be the case that there is no metaphysical entailment from the mass role, the charge role etc. to the consciousness-causing role.

This means that the zombie argument can succeed, even if dispositional essentialism is true – but only if the supervenience base is defined to exclude the consciousness-causing role. The zombie argument would then need to show that there is a possible world that is a minimal physical duplicate of the actual world, *minus any specific consciousness-causing disposition*, but which lacks consciousness. This, in turn, means that the definition of physicalism would need to be amended, so that physicalism is true if and only if the physical facts of the actual world, *minus any specific consciousness-causing disposition*, are metaphysically sufficient for consciousness. But this is just a terminological issue. The substantive metaphysical point – whether we are dispositional essentialists or not – is that there is no metaphysical entailment from the ordinary physical roles – the mass role, the charge role etc. – to consciousness.

5.5 - The Problem of Zombie Judgements

The inhabitants of the zombie universe really do look *exactly* like us from the outside – and this can have some very strange consequences (cf Chalmers 2010, p156). For example, in the zombie universe there is a zombie doppelganger of me, presently writing a thesis that is word-for-word identical to this one. My zombie counterpart would, if asked, deny that it is a zombie. But it would insist that zombies are metaphysically possible, and that this disproves physicalism. What follows? Should we think of my zombie counterpart as making any judgements at all, given that it is not conscious? Are its judgements correct? Is there some sort of incoherence here?

My own view is that while this scenario is strange, and seems paradoxical, it does not really present any serious philosophical difficulties. Whether or not we regard my zombie twin as making judgements depends on whether or not we think that intentional mental states can be given a reductive functional analysis. If they can be reduced in this way, then we may well take the view that my counterpart makes exactly the same judgements as I do. If we think that subjective consciousness is necessary for the possession of intentional mental states, then we will not think that my twin is making any judgements; it only seems to be. My own view is that there is an ambiguity in many of our intentional concepts. There is a sense in which states such as belief and judgement can be given a reductive functional analysis, and can exist without subjective consciousness; but there is also another sense in which they cannot be reductively analysed, and can only be possessed by conscious entities. The concept of representation is itself plausibly ambiguous, and can refer either to conscious representation, or to merely functional representational states of the sort that a non-conscious entity such as a computer might have. So my zombie twin makes the same functional judgements as me; but he does not make any conscious judgements.

But if we are willing to allow that there is a sense in which the zombie is making judgements, the next question is: are its judgements correct? Again, I do not think there is any fundamental problem here. My zombie twin's judgement that it is not a zombie is false. But its judgement that zombies are metaphysically possible, and that physicalism is therefore false, is correct. There is no fundamental difference between the zombie case and the case of a machine or robot in our own world that is functionally similar to a human, but which is not conscious. Such a machine – call it a *judgement machine* – might conceivably make the (functionally defined) judgements that it is not itself a judgement machine, but that judgement machines are possible. We ought to conclude that it is wrong on the first count, but right on the second. And whilst this would be a strange situation, it would not be paradoxical or incoherent.

Another version of this objection (e.g. Balog 1999) is the *zombie parity* objection.[28] The argument is that my zombie twin, following exactly the same reasoning as me, will come to the same conclusion – that physicalism is false. But – so the argument goes – physicalism is *true* in the zombie world. So there must be something wrong with the two-dimensional Argument. The problem with this objection is that physicalism is *false* in the zombie world. This might seem counter-intuitive – surely physicalism is true in the zombie world? But I do not think this is right. If physicalism were merely the claim that there is only physical stuff, then it would clearly be true in the zombie world. But this is not all there is to physicalism. What may be true is that physicalism is partly motivated by the belief that, in the actual world, there is only physical stuff. But physicalism is not just the claim, of a world, that it contains only physical stuff: it is the claim that the physical properties of the actual world are metaphysically sufficient for consciousness. Thus physicalism is a claim about the space of possible worlds: it is the thesis that there is no possible world that is a minimal physical duplicate of the actual world that fails to duplicate the consciousness of the actual world. But this claim is false, and it is false in all possible worlds – including the zombie world. Moreover, the physical properties of the zombie world are – by definition – not metaphysically sufficient for consciousness. So physicalism is false in the zombie world – and my zombie twin, to the extent that he makes any judgements at all, would be quite correct to endorse the zombie argument.

Conclusion

I began this thesis with the question whether or not consciousness is a purely physical phenomenon. The two-dimensional argument against physicalism shows that it is not. Physicalism entails that there is no possible world that is a minimal physical duplicate of the actual world, but which differs with respect to consciousness. But if, as I have argued, philosophical zombies are ideally conceivable, and if, as I have also argued, modal monism is

[28] It may be objected that my treatment of Balog's objection does not get to the heart of the challenge, which is that the zombie's pain* concept refers to c-fibres firing, and hence that its statement 'pain* is c-fibres firing' is a strong necessity. But I do not think that this is a strong challenge. The zombie's statement 'there is something it is like to be me' is demonstrably false. But its statements 'I am conscious' and 'I have consciousness' mean the same thing, and so are also false. But if its expressions 'my consciousness' or 'my pain' referred to things such as c-fibre stimulation, then these statements would be true.

true, then it follows that there is a metaphysically possible zombie world, and that physicalism is false.

Therefore some form of property dualism (at minimum) must be true. This means that mental property types associated with consciousness are not identical to physical property types – and, just as importantly, the facts about consciousness do not metaphysically supervene on the total distribution of physical properties in the universe. This does not rule out the causal supervenience of consciousness on the physical facts – indeed, I think that this probably holds in the actual world, although it is an empirical question. But causal supervenience does require the existence of psycho-physical connecting laws that are not themselves part of the microphysics of the universe.

Dualism inevitably raises an issue that lies well beyond the scope of this thesis. This concerns how consciousness is causally connected to the physical world, if conscious properties are not themselves physical and are not – as opposed to those found in the physical special sciences – merely functional properties of physical systems. This issue is closely related to a common objection to dualism, which is that dualism plus the causal closure of the physical realm entails the epiphenomenalism of the mental realm. The related causal argument for physicalism (e.g. Papineau 2002) is a perhaps the most important argument in favour of physicalism. My own view is that it is possible for dualism to overcome these objections. But that is an argument for another time.

The focus of this thesis so far has been on the problem of consciousness and the argument against physicalism. In the final two chapters, I will turn to the related issue of mental content, and whether or not it essentially depends on objects in the external world.

Chapter 6: Ain’t Meanings in the Head? (Part I)

Introduction

Externalism is the claim that the representational contents of our thoughts depend – not just causally, but essentially – on objects in the external world. It is summed up in Putnam’s famous dictum that *meanings just ain’t in the head* (Putnam, 1973, p 704).

The motivation for externalism comes from several well-known thought-experiments. These have a common structure: they invite us to imagine an individual with a particular mental content, and then suppose that his external circumstances (or those of a qualitatively identical duplicate) are varied in some relevant way, whilst all the internal facts are held constant. Crucially, everything seems the same from the subjective point of view of the characters involved, despite the change in their external circumstances. The resulting scenario is supposed to elicit from us the verdict that the mental content has changed in response to the changing external circumstances, even though the internal facts, including the phenomenal ones, remain constant. The most influential of these thought-experiments include *Twin Earth* (Putnam, 1973) and *Arthritis* (Burge, 1979). It is widely thought that these thought-experiments show that at least some of our mental contents depend essentially on objects in the external world.

The aim of this chapter is two-fold. The first aim is to show – contra the widespread view – that *Twin Earth*, *Arthritis*, and similar thought-experiments do not prove the case for externalism about mental content. I will show that it is possible to give an internalist interpretation of these cases, using two-dimensional semantics. The goal here is defensive: to show that the standard externalist thought-experiments do not prove the case for any strong externalist thesis about mental content. The second aim is to show that certain considerations about how mental content is grounded – known as Phenomenal Intentionality Theories (PIT) provide independent grounds for thinking that strong externalism about mental content is false.

In Section 6.1, I will define externalism more precisely: one can be an externalist about linguistic content without necessarily endorsing externalism about mental content; and there are strong and weak versions of mental externalism. In 6.2, I will outline the Twin Earth and Arthritis thought-experiments, and some initial lessons that can be drawn from them. Then in

6.3, I will set out a two-dimensional analysis of mental contents that permits an internalist reading of Twin Earth and similar thought-experiments; hence they do not prove the case for any strong version of mental externalism.

In Section 6.4, I will outline how mental externalism is incompatible with so-called Phenomenal Intentionality Theories (PIT), according to which mental content is constituted by phenomenal consciousness. Therefore, if we have independent reason to think that some form of PIT is true, this will be a reason to think externalism is false. Finally, in Section 6.5, I will set out an argument based rule-following for thinking that some form of PIT must be true – and therefore that externalism is false.

6.1 - What is Externalism?

What does it mean to say that meanings ain't in the head? There are two ways to interpret Putnam's dictum. The first is the literal interpretation: that mental contents essentially depend on objects that are outside the skin. But there is a second, deeper meaning, which is that mental contents essentially depend upon the existence in the world of the things that they are *about.*

To see the difference, suppose that when we think about our own brain states, our thought-contents essentially depend on what is actually going on inside our heads. If we have one sort of brain state, then our thought-contents will be fixed accordingly; but if we have a different sort of brain state, then our thoughts will vary accordingly, even if they seem subjectively the same. Does this amount to externalism? According to the first, literal definition, it does not: our brain states are inside our heads, and hence not external. But according to the deeper interpretation, it *is* a form of externalism: the contents of our thoughts depend on the real existence of the intentional objects of thought. It does not really matter that the object of thought in this case happen to be located inside the thinker's own skull. What matters is that the content of thought depends on the actual thing we are thinking about.

Thus externalism can be interpreted either in literal terms, or more deeply, as a claim about the relationship between our thoughts and the things they are *about*.[29] The deeper meaning of externalism becomes apparent when we consider what is required for a subject to be thinking *about* an object. Let us suppose that a thinker, T, is thinking about some object, O. How should we make sense of this? There seem to be two ways of construing what is going on. The first is that there exists some object, O, and that T stands in some appropriate relationship to it, such that T represents O. This way of construing matters entails a commitment to the actual existence of O, whatever it may be. We might term this construal, which entails the existence of the thing T is thinking about, *referential* aboutness. But there is another way of construing things, which is that T's thinking about O just consists in T entertaining a mental content that there exists an object, O. This does not seem to commit us to the actual existence of an object, O, to which T is related. Maybe O exists, maybe it doesn't – that is a further matter. We might term this way of construing things, in which we are not committed to the actual existence of O, *intentional* aboutness. So, we get something like the following distinction:

Referential aboutness (for thinker, T, and object, O):

$$\exists(x)\ \exists(y)\ [(x \text{ is } T)\ \&\ (y \text{ is } O)\ \&\ (x \text{ represents } y)]$$

Intentional aboutness (for thinker, T, and object, O):

$$\exists(x)\ [(x \text{ is } T)\ \&\ \{x \text{ represents that: } \exists(y)\ (y \text{ is } O)\}]$$

[29] Farkas (2008) offers an alternative definition of the internalism / externalism debate. She defines internalism as the thesis that a subject's consciousness determines all of their mental properties, and externalism as the denial of this. However, I will not adopt this definition for two reasons. First, I do not claim that consciousness determines *all* mental properties: it seems quite plausible to me that we may have mental properties that are not fixed by consciousness, such as standing propositional attitudes, and broad, secondary mental contents. Second, I think that Farkas's definition of internalism assumes too much at the outset. I understand externalism as the thesis that mental contents essentially depend on external conditions; internalism is therefore the denial of that claim. If internalism is true, then it follows that mental contents are internally constituted. But it does not obviously follow that it is constituted by *consciousness*. In fact, I do think that primary mental content is constituted by consciousness (a version of the Phenomenal Intentionality Theory) – but this requires further argument, in my view.

These two types of aboutness have quite different logical structures. In referential aboutness, we are quantifying over both the thinker and the object of thought, and asserting that they are related in such a way that one represents the other. In intentional aboutness, we are only quantifying over the thinker; the quantification over the object of thought occurs only within the scope of an intentionality operator. *We* are not asserting that the object, O, exists – we are merely asserting that the thinker, T, so asserts.

We can understand the difference between internalism and externalism is in terms of how they treat these two kinds of aboutness. Specifically, the question is whether intentional aboutness can occur without referential aboutness. In simple terms: if internalism is true, then intentional aboutness never requires referential aboutness, because (primary) mental contents are internally constituted and are therefore independent of external objects. But if externalism is true, then there are *some* contents that essentially depend on external objects, and for which intentional aboutness therefore depends on referential aboutness.

(Of course, even externalists agree that we can think about things that do not exist, such as dragons. But this raises the question of where such concepts come from if externalism is true. Presumably, externalism does not apply to our dragon-concepts – the content of our concept does not depend on the reality of dragons. Rather, our dragon-concept is composed or otherwise constructed from more mundane concepts: concerning reptiles, large size, the ability to fly, and so forth – all of which refer to real external objects. So externalism does not claim to apply to *all* external-world concepts. Even if externalism is true, there may be instances of intentional aboutness that are not referential. Rather, the point of externalism is that our external-world conceptual scheme has to be grounded in a relation to real external objects. This is what internalism denies.)

A further complication is that there are different versions of the externalist thesis. It is important to distinguish between the claim that some linguistic contents are externally constituted, and the claim that some mental contents are. And it is important to distinguish between the claim that some mental states have additional, secondary contents that are externally constituted, and the claim that some mental states *only* have external content.

I will not argue against linguistic externalism in this thesis. Indeed, I regard it as a truism that some *linguistic* content is constituted by external objects – words, after all, are public objects whose meanings are determined by facts beyond the internal constitution of the speaker. And I am open to the idea that at least some *mental* content is external. After all, once it is agreed that some linguistic content is external, it is not clear that it is possible to insulate all mental content from externalism. We usually frame our own thoughts using public language. And if I entertain a thought that is constructed using words with external content, what follows? Does my thought have an associated external content? Or should we distinguish between the linguistic content associated with my thought, which is externally constituted, and the mental content, which is a function of my consciousness? I will not take a stand on this issue. I am open to the idea that some thoughts have an associated external content, and I am agnostic as to whether we should assign such content to the mental or linguistic realm. But if it is assigned to the mental realm, then I will argue that it will only be as an additional, non-essential content, over and above the essential, internal one. What I reject, though, is the stronger claim that there are some thoughts whose *only* content is external. I will argue that, for any thought, it is always possible to define an internally constituted content, using two-dimensional semantics. This means that all thoughts, whatever external content they may have associated with them, also have an internally constituted truth-conditional content – a meaning in the head, so to speak.

In summary, we can identify three different versions of externalism (cf McGinn, 1989):

> a) Linguistic Externalism: Many linguistic items have semantic content that depends on factors external to the speaker.
>
> b) Weak Mental Externalism: Some thoughts have associated secondary and non-essential mental content that is constituted by external factors.
>
> c) Strong Mental Externalism: Some thoughts are such that they *only* have external content. They are constituted by an external relation.

So, of these versions of externalism, I accept (a) as a truism; I am agnostic with respect to (b) (whether (b) follows automatically from (a) depends on exactly how we should draw the boundary between what is considered linguistic content and what counts as mental content); but I reject (c). And it is the strong version of externalism, (c), that is the target of my argument in this chapter and the next. Whenever I refer to externalism *simpliciter*, without further qualification or specification, it is the strong mental version that I am talking about, and which is the target of my argument.

6.2 – The Externalist Thought-Experiments

The question before us is whether the classic externalist thought-experiments provide sufficient justification for strong externalism about mental content. I will argue that they do not.

Twin Earth invites us to imagine Oscar, an inhabitant of our own planet who, in the normal course of life, has various thoughts about water. We are then asked to imagine Twin Earth, existing in a distant part of the universe.[30] Twin Earth is molecule-for-molecule qualitatively identical to our own Earth, except that in place of water there is another substance with a different chemical formula, abbreviated to XYZ, but which nevertheless has all the same observable physical properties, and which performs exactly the same role as water in our own world. Things seem exactly the same to Twin Oscar as they do to Oscar; their situations are subjectively indiscriminable. So Oscar and Twin Oscar ('Toscar') have qualitatively identical psychological states. (This of course means we have to suppose that both are ignorant of the chemistry of their respective worlds). Moreover, their internal physical facts – brain states and so forth – are qualitatively identical, except for the trivial difference that Oscar has H_2O in his

[30] In *Twin Earth*, the two planets are co-located in the same physical universe. We can also have a variation of the thought-experiment in which there are two possible worlds: the actual world, in which the watery substance is H_2O, and a counterfactual world, in which it is XYZ. Varying the thought-experiment in this way does not raise any substantial philosophical issues, but the two versions do need to be handled slightly differently. For a start, intensions are usually understood as functions from possible worlds to objects. But in the original version, Oscar and Toscar inhabit the same possible world. Hence we need to refine how we understand their respective intensions: in this case, they should be understood as functions from a *centred* possible world – that is, a world with a particular individual identified as a point of reference – to objects. So in the original case we have functions not from different worlds, but from different centres of the same world. Another difference is that in a further development of the thought-experiment, it is useful to imagine Oscar transported from Earth to Twin Earth, and to see how he interacts with his new situation. But of course this is only possible if both planets are part of the same world. So there is no substantial problem with treating Earth and Twin Earth as different possible worlds as opposed to just different planets – in fact, it makes things a bit simpler – but it does allow for less flexibility with the thought-experiment.

cells and Toscar has XYZ. Twin Oscar, in the normal course of events, has thoughts about twin-water (which of course he calls ‘water’). And here is the point: despite that fact that everything is the same internally, Oscar is thinking about water and Toscar is not. Twin Oscar is thinking about some other substance, XYZ. Their thoughts are about different things, and therefore have different contents – purely because of the difference in their external circumstances. Or at least, so we are supposed to think.

Arthritis invites us to imagine an individual – call him Arthur – who has a pain in his thigh. Arthur lives in our own English-speaking community, and (wrongly) thinks that such a pain is referred to as ‘arthritis’. Meanwhile, Twin Arthur, being internally identical to Arthur, also has a pain in his thigh and also thinks it is called ‘arthritis’. But it so happens that Twin Arthur is right, for he lives in a community of pseudo-English speakers, in which ‘arthritis’ is a correct term for any leg pain. So when both Arthur and Twin Arthur utter the sentence ‘I have arthritis’, one of them speaks the truth and the other speaks falsely. But the relevant internal facts – including the facts about their respective legs – are the same. So if it is not the relevant facts, it must be the truth-conditions of the two utterances that differ. But if the truth conditions are different, then the contents must differ. So Arthur and Twin Arthur, despite being identical internally, mean different things, and think different thoughts, by the sentence ‘I have arthritis’ – purely because of their different external circumstances

What should we make of these thought experiments? I believe there are three relatively straightforward lessons to be drawn, which are applicable to both experiments (and indeed to externalist intuitions generally):

The first lesson is that these thought experiments offer good evidence for externalism with respect to linguistic content. It is a truism that the meanings of words can vary with circumstances that are external to the consciousness of individual users, since words are public objects. Their meanings are determined at least in part by a community of use, which may include customs of deference to experts. Thus, in the case of *arthritis*, the content of a term can track expert opinion on scientific facts, irrespective of what Arthur thinks of the matter. Hence the words Arthur uses may not mean what Arthur thinks they mean. Moreover, it may be part of the practice of a linguistic community that a word should rigidly designate the particular object or substance that plays a certain role in the actual world, even though science has not yet determined the identity of that object or substance. Thus words can have contents that are external not just to a particular individual on a particular occasion – they may be external to all speakers of a language. So the *Arthritis* and *Twin Earth* scenarios go hand-in-hand to show that externalism is true of at least some linguistic content: the former because the facts about a community of use are external to me; the latter because linguistic contents can track natural facts, which may be external to the linguistic community as a whole.

But this brings us to the second lesson: there is no straightforward inference from linguistic externalism to strong externalism about mental content (cf Segal, 1999). Does externalism about thoughts automatically follow from externalism about words? At first sight, it might appear so, given that we often formulate our thoughts – internally, in our own heads – using language. If the words in which I formulate my thoughts have an external content, then what follows? Does my thought therefore have an external content? But it is not clear that this follows. There seems to be a distinction between what my thought means to me, that is, how I understand things, and the public meaning of the words in which I formulate that thought. Thus we need to draw a distinction between what my words themselves mean when considered as public objects, and what I mean by my words. The distinction seems equally applicable whether I say my words out loud, or formulate them silently to myself, in my head. So perhaps if I, being in the same situation as Arthur, think the thought ‘I have arthritis’, we should draw a distinction between the content of my thought considered as a construction in public English – that is, its linguistic content – and the content it has for me – that is, my mental content.

Perhaps this distinction is somewhat artificial, and there is no way to isolate my mental content from the content of the words in which I formulate my thoughts, so that the external linguistic content automatically generates an external mental content. But even if we take this line, this only yields a very weak version of externalism about mental content: it means only that, associated with my thought about water and arthritis, there will be some mental content that is constituted by facts outside my consciousness. It does nothing to rule out the additional existence of an internal mental content that is constituted only by how things seem to me.

And in an important sense, both versions of Arthur do have the same mental content: both believe that they have a pain in their respective legs, and both believe that ‘arthritis’ is the correct term for the pain. In one linguistic community, the speaker happens to be right in his choice of words; in the other, he happens to be wrong. The *sentence* ‘I have arthritis in my leg’ will have different truth-conditions, and therefore a different meaning, depending on the community in which it is uttered. It is no great surprise that the meanings of words are fixed at least in part by community standards, and that individual speakers can be in error about them. But the *thought* ‘I have arthritis in my leg’ means the same thing to both speakers – each of them thinks that they have a pain in their thigh, and that this pain is correctly labelled ‘arthritis’’. We can give a similar account of the two versions of Oscar. So there is no straightforward inference from externalism about linguistic content to externalism about mental content – and certainly not to any strong version of that claim.

The third lesson is that we cannot cheat our way to externalism just by describing and reporting thought-contents in terms of their external relations. Mental contents such as beliefs and intentions can be reported or described in terms of their relation to external objects, but this need not entail that the content in itself depends on the existence of those objects. At face value, there is a distinction between *de re* and *de dicto* ascriptions of intentional states. *De re* ascriptions pick out an object in the world and report how that object is related to a particular intentional state; *de dicto* ascriptions, on the other hand, describe things from the point of view of the subject – they relate what the subject thinks about the world and its contents, irrespective of what the world and its contents are actually like. This is illustrated by the masked man case: suppose that I encounter the terrifying man in the iron mask who, unbeknownst to me, is in fact my father. I am not normally scared of my father. After this encounter, is it true to say that I am scared of my father? In the *de re* mode of reporting my intentional states, it is true to say that I am: there is an object such that it is my father and I am scared of it. But in the *de dicto* sense, it is not true that I am scared of my father – that is not how the world seem to me, and to claim that I am would be to misrepresent my thoughts.

This distinction is reflected in how the different forms of ascription are expressed in terms of quantification. In *de re* ascriptions the existential quantifier has scope over the operator that introduces the belief; these correspond to the referential sense of aboutness. In *de dicto* ascriptions, the existential quantifier occurs within the scope of an operator which creates an intensional context; these correspond to the intentional sense of aboutness. So *de re* ascriptions have the following form:

$$\exists x \text{ [P Believes: (x is F)]}$$

And *de dicto* ascriptions take the form:

P Believes: $[\exists x\ (x \text{ is } F)]$

Externalism is trivially true of *de re* ascriptions of mental content. It is a truism that, in *de re* terms, Oscar's thoughts are about H_2O and Twin Oscar's are about XYZ. But this does not entail that their respective thoughts, as they understand them themselves – that is, individuated in *de dicto* terms – are different. Indeed, we might think that Oscar and Twin Oscar have the same *de dicto* thoughts precisely because things seem the same to each of them. They may in fact be confronted with different external facts, but those external facts are presented in the same, watery way to both of them. And, so the internalist view goes, it is the subjective mode of presentation that is relevant to their *de dicto* thoughts and therefore their mental content.

In this spirit, we might regard *Twin Earth* and *Arthritis* as inverted versions of Frege's Puzzle. In the Frege case, there is one object, Venus, which has two modes of presentation. Thus two senses may in fact refer to the same object. In the Twin Earth case, there are two objects, H_2O and XYZ, which have the same, watery mode of presentation. Thus it is possible for one sense – the 'water' thought shared by Oscar and Twin Oscar to refer to different objects in different circumstances. When we report the beliefs of Oscar and Twin Oscar, we can do so either in terms of the object their beliefs are about – which will lead to broad ascriptions of content – or in term of the mode of presentation of that object – which will lead to narrow belief ascriptions. The internalist will then argue that mental content itself is individuated in terms of the mode of presentation of the object, not the identity of the object itself, and is therefore the same for Oscar and his Twin, and located in their respective heads.

There is a deep worry for this sort of internalist account of meaning: what fixes the contents of the predicates that go to make up a mode of presentation? What fixes the meaning of *F* in the example formula above? There seems to be a dilemma. If such predicates get their contents from a relation to external objects, then the resulting account of meaning is actually externalist – and we run the risk that Oscar's *F* and Twin Oscar's *F* will differ in meaning. But if they do not, then the meaning of such predicates can seemingly only fixed by further descriptive contents – an object, x, is F if and only if it is both G and H, and so forth. But then what fixes the meanings of G and H? And so on, *ad infinitum*. For an account of meaning to be truly internalist, contents can never be cashed out in terms of actual external objects. But then it seems that all we are left with is an endless – or perhaps circular – set of descriptions, interpreting one another. How can anything in this closed circle mean anything in terms of objects external to the circle? And how does any such system manage to refer to objects in the world at all? The worry is that, if contents are not essentially connected to external objects, then reference to them becomes utterly mysterious. One option, which I will expand on in Chapter 7, Sections 7.2 to 7.4, is that (some at least) of our external world concepts have a primary intension that picks out whatever object is the systematic, external cause of certain experiences.

So it is not immediately obvious that *Twin Earth*, *Arthritis*, or other such thought-experiments entail any strong version of externalism about mental content. It is possible for an internalist to resist that conclusion (at least, for all we have seen so far), by invoking, first, the distinction between linguistic and mental content, and second, the distinction between *de re* and *de dicto* reports of intentional states. Perhaps the externalist thought-experiments do show that some linguistic contents depend on the external facts, and that mental contents are individuated in terms of external objects when they are reported in *de re* terms – but neither of these claims constitutes reason to endorse strong externalism about mental content.

6.3 – Twin Earth and Two-Dimensional Mental Content

In this section, I will show that it is possible to give an internalist account of the Twin Earth thought-experiment, using the framework of two-dimensional semantics. The ideas set out here will work equally well for *Arthritis*, and indeed any similar thought-experiment. The essential points are these:

a) Oscar and Twin Oscar share the same 1-intension for their respective ‘water’ concepts.

Although:

b) They have different 2-intensions for their respective ‘water’ concepts: Oscar’s 2-intension for ‘water’ is H_2O; Twin Oscar’s is XYZ.

However:

c) There is a mental content determined by the 1-intensions, and which is therefore the same in both cases, regardless of the external facts.

In order for the internalist interpretation to work, there must be a 1-intension that assigns Oscar's water-thoughts the same truth-conditions as Twin Oscar's corresponding water-thoughts. If they have different truth-conditions, then they will have different contents – and that is not what the internalist wants. Thus the best externalist line of attack is to argue that their respective water-thoughts have different truth conditions, and therefore there is no common content.

This point can be made by varying the thought-experiment as follows (cf Boghossian 1989a): suppose that Oscar is magically transported to Twin Earth, waking up there one morning, unaware of his new situation. He goes to the well to draw a bucket of the colourless liquid inside and thinks 'there is water'. But this thought is false – there is no water in the well, there is only XYZ. Yet when Twin-Oscar goes to a qualitatively identical well and thinks the subjectively indiscriminable thought 'there is water', his thought is true. So, confronted with the same set of external facts, their thoughts have different truth-values. Therefore they must have different truth conditions; and therefore they must have different contents. But, since Oscar and Twin-Oscar are qualitatively identical with respect to their internal facts, the difference in their mental contents must be due to their different relationship to the external circumstances.

The most straightforward way for an internalist to respond is to simply deny that Oscar's thought on Twin Earth that 'this is water' is false, when interpreted in terms of the 1-intension of his 'water' concept. The important point here is that we should not evaluate Oscar's thought in terms of *our* concept of water, but his. *Our* concept of water is very plausibly such that if we were transported to Twin Earth in the manner of Oscar, then we would be wrong to judge the stuff in the wells and lakes to be water. But then we know chemistry; it is plausibly part of our concept of water that it is H_2O. Perhaps Oscar's concept is different. It is at least plausible that Oscar's (and indeed Twin Oscar's) concept of water is just a descriptive function roughly equivalent to 'any stuff with the watery properties'. In this case, anything that has the relevant observable physical properties counts as being water according to their concept of it, irrespective of its microphysical structure. In that case, both Oscar and Twin Oscar would be right to identify the watery stuff on Twin Earth as water; and they would both be right to do so on Earth, as well. If their concept of water is just a simple descriptive function along these lines, then, upon learning chemistry and being told that the stuff on Earth is H2O and the stuff on Twin Earth is XYZ, they would conclude that on Earth water is H2O, and on Twin Earth water is XYZ.

So if natural kind concepts such as ‘water’ are simple descriptive functions like ‘watery substance’ then their contents do not depend upon the external physical facts. And of course it is possible to define a concept ‘water’ that functions in just this way. But that seems like too easy a victory for internalism, since that is not how our water-concept actually works. We would not take the concept ‘water’ to denote just any watery stuff: it would have to be *water* – which, as chemists now know, is H_2O. If we were to discover, somewhere on Earth, a watery substance that, upon chemical analysis, turned out to be XYZ, we would not classify it as *water*. We would regard it as something else entirely – *fool’s water,* perhaps (cf Kripke 1980).

So if Oscar’s water concept functions like ours, then it should not pick out just any watery substance, irrespective of its microphysical structure. On learning chemistry, and being told that the stuff on Earth is H_2O, and the stuff on Twin Earth is XYZ, he should conclude that he was mistaken to call the stuff on Twin Earth ‘water’. He should take the view that it looked like water, it seemed like water – but it wasn’t really water at all. On learning all of this, he should make clear that, by the concept ‘water’, he always understood *that particular watery stuff that he interacted with on Earth* – and not just any old stuff which has the same observable physical properties. But doesn’t this mean that the content of the concept depends on the external circumstances? Does it not entail that the content of ‘water’ on Earth is H_2O, and on Twin Earth it is XYZ? The challenge for internalism is to explain how Oscar and Twin Oscar could have the same mental content if their respective ‘water’ concepts function in this way.

So internalism should reject the easy path. It is inadequate to regard natural kind concepts such as ‘water’ as simple descriptive functions, which are insensitive to the microphysical facts. We have to concede to the externalist the point that, as it happens, our concepts are indeed sensitive to external facts. Therefore a successful internalist account of mental content must show that Oscar and Twin Oscar have a common 1-intension – but that this intension yields different judgements about what counts as water in their respective cases. So it needs to show how their respective thoughts have the same truth-conditions, but, because they are confronted by different sets of facts, their thoughts can have different truth-values.

How can internalism meet this challenge? The key is to understand the relationship between content, truth conditions, and reference. Content is determined by truth-conditions. Thus the content of a statement just is its truth-conditions; where the statement is ambiguous, and may be interpreted as expressing a number of different propositions, each of these will correspond to a different truth-condition. Similarly, although we do not evaluate referring expressions themselves as being true or false, the content of a referring expression is determined by the conditions under which it is satisfied – that is, the conditions under which it picks out its reference. Therefore, where an expression has different primary and secondary intensions, we are liable to find that there are different satisfaction-conditions, and these will determine different contents.

What is the relationship between content and reference? Some Fregeans (e.g. Searle 1983), and maybe Frege himself, thought that sense determines reference: once we know the sense of an expression, we know the content; and once we know the content, we know the reference. But this will not always be true on my proposal. On my view, the sense and the content are the same thing, because the meaning is just the truth (or satisfaction) conditions. But content will not always determine reference – at least, not on its own. Whether content determines reference will depend on what type of intension we have. For 2-intensions, the content is identical to the reference. Thus the 2-intension of 'water' will be H_2O on Earth, and XYZ on Twin Earth. So externalism is trivially true of secondary contents. But what about 1-intensions?

For 1-intensions, the picture is more complicated. For 1-intensions that consist of a simple descriptive function – such as 'watery stuff' – the content will not be identical to the reference, but it will be sufficient to determine the reference. The reference will just be whatever satisfies the description. So the 1-intension of 'watery stuff' determines the reference as being H_2O on Earth, and XYZ on Twin Earth. So far, this is relatively straightforward. But the situation is more complicated when we consider 1-intensions that consist of comparative – that is, two-place – descriptive functions. In order for an object to satisfy a two-place function, it is not sufficient for the object to display the relevant properties (though this is likely to be a necessary condition). It must also be the same object that displayed the relevant properties on some previous occasion or occasions. A 1-intension for 'water' along these lines would be a description of the form: 'the watery stuff I previously encountered'. If this is the 1-intension for 'water', then it will only be satisfied by the substance now in front of me if that substance has all the relevant watery properties and is the *same* microphysical substance that I am used to. So, for 1-intensions that have a two-place or comparative structure, reference is not determined by content alone, but by content plus some contextual facts, typically those relating to an initial baptism or use of the referring expression. So, although the content of a referring expression is determined by its satisfaction conditions – that is, the conditions under which it picks out its reference – this is consistent with the same content picking out different references under different conditions.

Now, what is the primary content of the judgement ‘this is water’? If we are judging just that it is the same watery stuff as we previously encountered, then the truth of the judgement – and the content of the concept ‘water’ – does not depend upon *which* stuff it is. It doesn’t matter whether it is H_2O or XYZ, as long as it is the same in both instances. The judgement is comparative, not identity dependent. So, applying this model to the Twin Earth case, we will find that both Oscar and Twin Oscar have the same concept of water. Each, on visiting the well, will think (roughly): ‘There is the same watery substance that I previously encountered’. Both individuals have the same content, and therefore their thoughts must have the same truth-conditions. But, because of the difference between their original contexts of use for the term ‘water’, the facts confronting Oscar are different to those confronting Twin Oscar. So Oscar’s judgement is false, and Twin Oscar’s identical judgement is true.[31]

We can formalise this way of analysing the 1-intension of the sentence ‘this is water’ in the following way. A simple one-place descriptive function – which I previously said is inadequate – would have something like the following structure:

$$\exists x\ [(\text{Is watery})\ x\ \&\ (\text{I am now indicating})\ x]$$

But if we treat the 1-intension as a comparative function, we get something like the following:

[31] This account also has the advantage that it allows for the reference of a concept to change over time, as the thinker’s circumstances and experience change. Thus Boghossian (1989) argues that although Oscar is wrong to call XYZ ‘water’ when he *first* arrives on Twin Earth, this usage will become correct for him after a period of adjustment. By interacting with XYZ for long enough, it will become ‘the watery stuff that [he] is used to interacting with in [his] local environment’, and thus will come to satisfy his concept of water.

> ∃x ∃y {[(Is watery) x & (I was previously habituated to) x] & [(Is watery) y & (I am now indicating) y] & (x = y)}

In the first case, I am indicating a substance in a present circumstance of evaluation and attributing to it the relevant, watery properties. In the second case, I am doing all this, and more: I am also stating that the watery substance in the present circumstance of evaluation is the same microphysical substance as the watery stuff in an earlier (or habitual) context of use.

There are obvious parallels between this two-dimensional account and Kaplan's distinction between the *character* of a referring expression (analogous to the 1-intension), and its *content* (analogous to the 2-intension), which I discussed in Chapter 3, Section 3.1. The crucial difference though, is that, for Kaplan, the character is not sufficient to determine truth-conditions. Character must be applied within a context of use to determine the content, which determines truth-conditions. Hence – according to this account – truth-conditions always depend in part upon the actual facts of the context of use – they are, so to speak, external. But in the two-dimensional account that I am advocating, 1-intensions *can* determine truth conditions independently of the facts of a particular context of use – hence they are not exactly analogous to Kaplan's notion of character.

It may be objected that my notion of the content of these comparative 1-intensions is inadequate – that it is not really content because, on its own, it does not determine reference. Would it not be better – so the objection goes – to regard the content as determined by the function plus the relevant contextual facts?

Of course, if this objection is right, then it would be fatal to my internalist account of Twin Earth. My proposal depends on the idea that Oscar and Twin Oscar can have the same 1-intensions for 'water' even though their respective intensions have different references, due to the different historical contexts of Oscar and Twin Oscar. But I do not think the objection is correct. It is a mistake, in my view, to equate content with reference. Content is about truth-conditions. That is, content determines a set of possible worlds in which a proposition is true, or in which a referring expression is satisfied. Where the content is a two-place function, this set of words will look somewhat disjunctive. If the 1-intension of water is indeed 'the watery stuff I am used to', then it will be satisfied in worlds where there is H_2O in both the original context of use and the current circumstance of evaluation, and in worlds where there is XYZ in both places – but not in worlds where there is H_2O in one place, but XYZ in the other. The important thing is that the 1-intension does determine a set of worlds in which it is satisfied – and therefore it is a semantic content. But it will be satisfied by different objects in different worlds – indeed, potentially for different agents within the same world – depending upon the history of use of the concept in that world.

I think a lot of our concepts work like this; the phenomenon is far more widespread than might at first sight appear. Concept-application is often not just about categorising an object on the basis of the properties it displays in a given situation. It is often about using those properties to determine that an object is the very same thing (the same individual, the same type) that has been previously encountered. What really counts in these cases is that the object should in fact be the same thing, by whatever the relevant standards of sameness are (microphysical structure, causal-historical continuity, and so forth).

The same idea will also work for proper names, provided that they can have a descriptive sense that is constituted by a recursive or two-place function. Suppose, for example, that Oscar has a friend (on Earth) called 'Paul'. Meanwhile, Twin-Oscar has a physically identical friend, who he also calls 'Paul' – but who we shall call 'Schmaul'. Once again, Oscar is magically transported to Twin-Earth. This time, he bumps in to Schmaul – and thinks 'there's Paul'. His thought is false. But when Twin-Oscar thinks 'there's Paul', his thought is true. So the counter-argument seems to go through here – it seems that the content of the Paul-thoughts is broad. My own view is that it is possible to construct an internalist counter-argument here as well. In outline, this involves a theory of proper names as meta-linguistic descriptions, and the content of Paul-thoughts being equivalent to (something like): 'there is the person who is a bearer of 'Paul' and it is the same bearer of 'Paul' that I met previously'. Thus Oscar and Twin-Oscar would have the same narrow thoughts when confronted by Schmaul, but Oscar's thought is false and Twin-Oscar's true because, although confronted by the same individual, they are confronted by different sets of facts – the crucial thing is that for Twin Oscar, Schmaul does count as being the same bearer of 'Paul' that he previously met, whereas for Oscar this is not the case.

So, in summary, it is possible to give an internalist reading of Twin Earth using the framework of two-dimensional semantics. It is perfectly conceivable that Toscar and Oscar have the same internally-constituted primary contents, despite their external differences. The idea is that

Oscar and Twin Oscar have the same 1-intension for ‘water’, but different 2-intensions. The 1-intensions are narrow – their content is independent of the external facts – whereas the 2-intensions are broad, and are constituted by the external facts. In order to account for the fact that Oscar’s water-thoughts are false on Twin Earth (at least initially), the 1-intension in this case would have to be a two-place function that relates an object in the circumstance of evaluation to an original context of use, and says that it is the same object. Oscar and Twin Oscar will therefore have the same primary truth conditions for their water-thoughts, namely that the watery stuff they are confronted with in a particular instance is the same watery stuff that they habitually interact with. But they will have different secondary truth conditions for their water-thoughts – Oscar’s will relate to H_2O, Twin Oscar’s to XYZ.

6.4 – Externalism and Phenomenal Intentionality

So far in this chapter, I have argued that Twin Earth and similar thought-experiments do not prove the case for strong externalism, because we can define primary intensions for Oscar and Toscar that are independent of their external facts, and which determine the same truth conditions for both. But this then raises the question: what determines our primary intensions? More specifically, what makes it true of a thinker that their thoughts have primary intensions at all? And what makes it true of any given thought that it has one particular primary intension, and not another? Up to this point, I have assumed that primary intensions themselves are independent of the external facts. But, unless we can give a plausible account of how this is so, the whole argument against externalism may seem question-begging. So the externalist could argue that, even if it is granted that primary intensions are sufficient to the determine truth-conditions of our thoughts independently of the external facts, there is no reason to suppose that primary intensions themselves are internally constituted.

What options are open to internalism in response? At face value, there seem to be several possibilities. One is that primary intensions are constituted by the functional roles that brain states play within our cognitive apparatus. But this is not my view. Much the most straightforward internalist account, which I endorse, is that primary intensions are a function of consciousness. This is known as the Phenomenal Intentionality Theory (PIT), as set out in Bourget & Mendelovici (2017). PIT is the claim that the facts about the intentional content of a certain class of mental items – in this case, conscious thoughts – is determined by their phenomenology.

PIT need not entail that all representational content, belonging to all kinds of state, is determined by consciousness. Whether it does so or not will depend upon the scope of the favoured version of the theory – that is, the classes of mental representation to which it applies. For my purposes, in support of the two-dimensional argument against externalism, it is only necessary that PIT should be true of occurrent, conscious, narrow mental contents. So there is no need to argue that phenomenal consciousness determines the representational content of standing propositional attitudes, or unconscious mental states, or broad, secondary contents. The weak version of phenomenal intentionality theory that I favour is perfectly consistent with standing propositional attitudes and unconscious beliefs being functionally constituted, and with externalism with respect to secondary mental contents, and I will not argue otherwise.[32]

Moreover, PIT need not be a reductive theory of mental content; it does not necessarily mean that representational content is determined by phenomenal consciousness that can itself be characterised in non-intentional terms. As Chalmers (2010) puts it:

> It is not implausible that there is something about consciousness that by its very nature yields representations of the world. One might hold that at least with perceptual experiences, representational content accrues *in virtue* of the phenomenology. One might hold that something similar holds for beliefs […]
>
> Still, for this approach to provide a reductive grounding for the intentional, we would need to characterize the underlying phenomenal domain in nonintentional terms, and it is far from clear that this is possible. [p371; author's italics]

In fact, I do not favour any such reductive version of PIT. My own view is that some phenomenal consciousness is intrinsically representative in character. There can be something it is like to have a particular understanding or intention; and, moreover, the semantic content in such cases is determined by the phenomenal character. We should not think of phenomenal

[32] One might make the further claim that standing propositional attitudes and unconscious mental states are dispositions to have occurrent, conscious mental states whose representational content is constituted by their phenomenology – but that is another matter.

consciousness just in sensory terms; I think there is also a distinctive cognitive or intellectual phenomenology.

Externalism is incompatible with PIT, because it entails that consciousness is not sufficient for mental content. Suppose that we interpret Putnam’s dictum literally, and we also make the reasonable assumption that consciousness supervenes causally on our internal physical states. In this case, externalism entails that two subjects can be identical with respect to their internal physical states (and therefore with respect to consciousness), but differ with respect to their contents. So this means that consciousness is not sufficient for content. Similarly, if we interpret the externalist premise in the deeper, non-literal sense, then we get the same result: mental contents depend on the actual objects of thought, which exist independently of our consciousness. So, if externalism is true, then the totality of facts about consciousness is not sufficient to determine the totality of facts about mental content. Thus externalism entails:

Q =/=> C

Where Q represents the totality of facts about consciousness, and C represents the totality of facts about mental content.

Therefore, if we have independent reason to think that mental contents must be constituted by consciousness, then this will in itself be a reason to think that externalism is false. But is there reason to think even this relatively limited version of PIT is true? To discuss this question in depth is beyond the scope of this thesis. But in this section and the next, I will offer two reasons to think that PIT is indeed true of our occurrent, conscious, primary intensions. The first arises from a modified version of the zombie argument; the second arises from certain considerations about rule following and meaning.

With respect to the first reason, the question arises as to what extent philosophical zombies can be said to have representational content. Plausibly, there is a perfectly intelligible sense in which they do, but also a sense in which they do not. What is certainly true is that a zombie’s internal physical states and its physical behaviour are identical to those of its non-zombie doppelganger. If it is true of me that I have internal physical states that track external objects, and systematically produce appropriate outputs (whether in the form of behaviour or other system inputs), then it is true of my zombie counterpart. Very plausibly, something along these lines provides a functional analysis of what it is to have representational states. To the extent that representational content can be explained by tracking theories (and reductive conceptual role theories), it can be had by zombies.

But it also seems plausible that there is an important sense in which I have representations and my zombie double does not. The zombie may indeed represent things in the same way that a computer or machine does; but unlike me, it does not consciously represent – it does not have occurrent *thoughts*. But the only difference between the actual world and the zombie world is the presence or otherwise of phenomenal consciousness. So if there is a sense in which I have representations that my zombie counterpart lacks, these can only be constituted by the phenomenal consciousness that is present in me but not in the zombie. If this is right, it shows that some version of PIT must be true, at least with respect to a certain notion of representation.

Now, it may be objected that my conscious representations are not constituted by my phenomenal consciousness alone, but by the conjunction of phenomenal consciousness with the representational functions instantiated by the underlying physics. But I do not think this objection will stand up. For there is a conceivable – and therefore metaphysically possible – world that is a duplicate of the actual world with respect to phenomenal consciousness, but which lacks any physical properties (in other words, there is a possible world in which the Cartesian hypothesis of an all-powerful deceiver is realised). And there is no *a priori* reason why this world should not have all of the same conscious thought-contents as the actual world. So if there is a possible world in which the Cartesian thought-experiment about an all-powerful deceiver turns out to be true, then phenomenal consciousness is metaphysically sufficient to constitute the representational content of occurrent, conscious thoughts.[33]

6.5 – Phenomenal Intentionality and the Rule-Following Problem

The second reason for thinking that some form of PIT must be true comes from certain considerations about meaning and rule-following. Kripke (1982) presents a sceptical problem

[33] Of course, this raises the problem of how such a disembodied mind could refer to worldly objects and properties – and how our own mental contents, if so constituted, refer to the world. I will return to this issue in Chapter 7.

about meanings, based on certain considerations about rule-following; Goff (2012) argues that the only way to overcome these problems and resist scepticism about meanings is to endorse some form of PIT. Goff's argument, in outline, is that there are facts about mental contents – but such facts require there to be facts about rule-following on the part of thinkers. However, facts about rule-following are under-determined by any facts other than the phenomenal facts; therefore facts about meaning can only be grounded in the phenomenal facts. In the remainder of this section, I will explain this argument in more detail and outline some of the further questions it raises.

In 'Wittgenstein on Rules and Private Language' (1982) Kripke presents his interpretation of some of the later Wittgenstein's ideas about language. The central thesis of this interpretation is that there are no facts about meaning. This – irony intended – means exactly what it seems to mean: that there is literally no fact such that I mean one specific thing, and not any other, by any given linguistic performance. Let us call this the Sceptical Thesis about meaning. Though the argument of Kripke's Wittgenstein ('Kripkenstein') is primarily directed at language, the same considerations apply equally well to mental contents.

The underlying premise of the argument is that a linguistic performance can only have a determinate meaning if the correctness or otherwise of the performance is determined by a rule. That is, there must be a rule in force which makes one particular performance the right one in the circumstances to convey the particular meaning that I want to convey. For example, if I want, in normal circumstances, to express the fact that two plus three equals five, then it would be correct for me to utter the English sentence 'Two plus three equals five', or something similar. But it would not be correct to utter the sentence 'Two times three equals six'. The latter sentence does, in fact, express a meaning in English – and it also happens to express a mathematical truth. But it is not the *same* meaning, and it is not the right one in the circumstances for what I want to say. It would be the wrong thing to say, not mathematically, but linguistically. Presumably then, there is a set of rules in force that connects my intended meaning with the circumstances in which I find myself, and determines the correct linguistic performance for me. If there were no such rule, then we could not evaluate my particular performance as being either correct or incorrect in the circumstances – and this would undermine the idea that it meant anything in particular. Furthermore, following a rule is not the same as merely conforming to a rule. As we shall see, there are very many rules to which my behaviour on a particular occasion might be interpreted as conforming. But this is not sufficient to determine which rule out of the many possible rules is the one that I am actually following, and which gives the act its meaning.

For the sake of argument, I will assume the truth of the underlying premise that linguistic meaning, if it exists, requires rule-following. This is fundamental to everything that follows. Given this assumption, Kripkenstein poses the question: in virtue of what is it true of me that I am following one rule and not any other in a particular instance? His answer is that there is no fact in virtue of which it is true of me – and therefore no such thing as following (as opposed to merely conforming to) a rule, and therefore no such thing as determinate meaning in language. Thus he writes:

> In s201 [of the *Philosophical Investigations*] Wittgenstein says 'this was our paradox: no course of action could be determined by a rule […] I will attempt to develop the paradox in question. [p 8]

It is important to note that the Sceptical Thesis is not primarily an epistemological problem. It is not just that I cannot *know* which rule I am or should be following (as when I misremember my past linguistic intentions, and am therefore mistaken about what I should do in order to follow them). It is that there is no fact at all.

Kripkenstein gives the example of the *plus* rule in mathematics [p 8]. Suppose that up until the present, I have never added together any two numbers of which one has been greater than 56. I am now asked to add 57 and 68. If I am following the *plus* rule, then I should give the answer '125'. But the sceptic then asks what makes it true of me that I have been following the *plus* rule as opposed to the alternative *quus* rule, which is like the *plus* rule, except that if one of the terms in the addition is greater than 56, then one must automatically give the answer '5'. So:

> The sceptic doubts whether any instructions I gave myself in the past compel […] the answer '125' rather than '5' […] perhaps when I used the term 'plus' in the *past,* I always meant quus…[p 13; author's italics]

This may seem bizarre. But what makes it wrong? Clearly, it is no use appealing to the fact that I *should* follow this *plus* rule as opposed to *quus* – this is obviously question-begging. As Kripkenstein puts it:

> An answer to the sceptic must satisfy two conditions. First, it must give an account of what fact it is […] that constitutes my meaning plus, not quus. But further […] it must show […] how I am justified in giving the answer ‘125’ to ’68 + 57’. [p 11]

So, if the sceptic is wrong, then there must be some fact about me in virtue of which it is true that I was following *plus* as opposed to *quus* when I did the previous sums, and such that it would now be wrong to demonstrate the deviant, *quus*-like behaviour. The argument for the Sceptical Thesis consists of considering, in turn, the various contenders for the role of the fact in virtue of which I am following the *plus* rule – and rejecting them. I will not explore them all in detail, but will outline each one.

The first consideration is that there is nothing about my past behaviour which makes it true of me that I have been following the *plus* rule as opposed to *quus* up until now. By the terms of the thought-experiment, all of my past behaviour is consistent with either interpretation. And, no matter how much behavioural data we might gather – so the argument goes – there will always be indefinitely many ways of interpreting it, and indefinitely many rules to which it conforms. This point is similar to Quine’s (1960) account of the radical translator. The lesson is that facts about rule-following are underdetermined by facts about behaviour. But whereas Quine effectively stopped looking for the fact which justifies one interpretation over the others, having failed to find it in behavioural accounts, Kripkenstein will go on to consider, and find wanting, various other possibilities. So:

> Another important rule of the game is that there are no limitations, in particular, no *behaviourist* limitations, on the facts that may be cited to answer the sceptic […] This feature of Wittgenstein contrasts […] with Quine’s discussion of the ‘indeterminacy of translation’. […] Quine […] is more than content to assume that only behavioural evidence is to be admitted into his discussion. [p 14; author’s italics]

I will not explore this point in more detail here, but for the sake of argument I will accept that this under-determination thesis is correct, and facts about behaviour are not in themselves sufficient to secure facts about meaning.

The next contender for the role of meaning-fixing fact is behavioural *dispositions*. But the fact that I have been following the *plus* rule cannot just be a matter of my dispositions to behave in *plus*-like as opposed to *quus*-like ways – or so Kripkenstein argues [pp 22-23]. The general objection to a dispositionalist account of rule-following is that whatever I am disposed to do will, by definition, be correct – which is the same as following no rule at all. Of course, dispositionalists have acknowledged and attempted to answer this objection (cf Boghossian, 1989). I cannot deal with these attempts at present, save to say that I do not find them successful, and will accept Kripkenstein's point for the sake of argument.

Having ruled out actual behaviour and behavioural dispositions, Kripke considers whether *intentions* are sufficient to determine facts about rule-following. Can it be the case that I ought to answer '125', and not '5', because in the past I *intended plus* and not *quus*?

Kripkenstein argues that there is no such fact about any of my mental states, such as my past intentions. For how do my past intentions discriminate between *plus* and *quus*? Perhaps in the past I attempted to set up a rule as follows: I performed an addition, adding, say, 5 and 6 and giving the answer '11', and then vowed that in future I would carry on like *that.* But this will not do, for the very thing in question is what it means to carry on like *that*. Perhaps in the past I was in fact lucky enough to anticipate the *quus*-like deviation that occurs when we get to 57, and consciously ruled it out. But what if the deviation were to occur at 58, or 59? And so on – I cannot rule out all possible deviant interpretations in this way. So there is nothing that took place in my mind, nothing I can introspect, that will make it true of me that I meant *plus*. So he writes:

> […] nothing in my mental history of past behaviour – not even what an omniscient God would know – could establish whether I meant plus or quus. [p 21]

Moreover, I cannot fix the *plus*-interpretation by appealing to a further rule, without raising the same problem for that rule in turn. Kripkenstein explains this point with the following example [pp 16-17]: I might attempt to explain that I did mean *plus* all along by explaining the *plus* rule in terms of *counting:* I intend to make a pile of marbles (say) equal in number to one of the terms of the addition, and another pile for the other term, and I push the two piles together, and I *count* the resulting pile – and that is what I mean by the *plus*-function. But, asks the sceptic, how do I know that I am *counting* and not *quounting,* where *quounting* is such that if one of the original piles is more than 56, then the resultant pile always gets the value '5'? The crucial point for Kripkenstein is that facts about meaning are not grounded in facts about past intentions because there is no subjective difference between intending plus and intending any of the infinitely many quus-like deviations that I have not specifically and consciously considered and ruled out. That is, there is nothing it is like to intend plus as opposed to quus – and, since there is no phenomenological difference, there is no way to ground the fact that I mean plus, and not quus, in some mental act of intending.

So we can summarise the argument for the Sceptical Thesis as follows:

> i) If there are facts about meanings, then there are facts about rule-following.
>
> But:
>
> ii) Facts about rule-following are underdetermined by facts about actual behaviour; nor are they determined by facts about behavioural dispositions.
>
> And:
>
> iii) Facts about rule-following are under-determined by facts about any mental states such as past intentions because they are under-determined by phenomenal consciousness: there is nothing it is like to intend plus, as opposed to quus.
>
> Further:
>
> iv) There is nothing else in virtue of which a fact about rule-following could be determined.
>
> Therefore:
>
> v) There is no fact in virtue of which it is true that a particular rule is being followed.
>
> Therefore:
>
> vi) There are no facts about meaning. [The Sceptical Thesis]

I will assume, for the sake of argument, that the conclusion of this argument is false, and that there are indeed facts about meanings. This, after all, is not in dispute between internalists and externalists. But if the Sceptical Thesis is false then Kripkenstein's argument must make a false step. If we can identify where it goes wrong, then this ought to tell us something important about how facts about meanings are grounded. The idea, then, is that identifying the false step will lead to an argument for some form of PIT.

Let us suppose, for the sake of argument, that we accept premise (i) of Kripkenstein's argument. That is, we accept that facts about meaning require there to be facts about which rules govern our linguistic behaviour and our thoughts – but we nonetheless insist that there are facts about meaning. Further, let us suppose for the sake of argument that such facts could only be grounded in either facts about behaviour, dispositions and so forth, or facts about the intrinsic properties of certain mental items – so we accept premise (v). Although these assumptions themselves are not uncontroversial, I will not argue for them here. But what is clear is that if these assumptions are right, then one of two things must be true: either Kripkenstein is wrong about the underdetermination of facts about rule-following by facts about behaviour or behavioural dispositions; or he is wrong to claim that no mental item is sufficient to constitute meanings. Therefore, if Kripkenstein is right that facts about rule-following are underdetermined by behaviour, then he must be wrong to deny that meanings are constituted by intrinsic phenomenal properties of intentions.

This means that the rule-following considerations can-be adapted to present an argument for PIT, along the following lines (Bourget & Mendelovici 2019, s4.5): there are facts about meanings; facts about meanings require the existence of facts about rule-following; facts about rule-following can only be constituted by either behavioural facts, or facts about the phenomenal properties of intentions; Kripkenstein is correct that rule-following is under-determined by behavioural facts; therefore Kripkenstein is wrong to deny that facts about rule-following are determined by facts about the phenomenal properties of intentions.[34]

[34] Goff (2012) employs the rule-following considerations, in conjunction with phenomenal intentionality, in order to present an argument against physicalism. He describes a variation on Jackson's (1982) knowledge argument, in which Mary tries to determine whether a subject means plus or quus by their utterances. Mary is armed with perfect knowledge of the subject's physical states (including brain states), behaviour, dispositions, and so forth. Goff argues, as per Kripkenstein, that Mary cannot know facts about the subject's mental contents on the basis of this knowledge. There is no *a priori* entailment from any of these basic facts to facts about meaning;

But there is an obvious objection to be made by opponents of phenomenal intentionality. The objection is that, as Kripkenstein observes, there is no subjective difference between intending *plus* and intending any of the infinitely many *quus*-like deviations that can also be made to fit my subsequent behaviour. So how can my phenomenology determine that I mean plus and not quus? This is a deep problem for any version of PIT. Our phenomenology is not infinitely fine-grained – and yet, to solve the rule-following problem, it seems that it needs to be. How can phenomenal consciousness determine representational content if there is nothing it is like to intend plus as opposed to quus?

I think it is possible for PIT to solve this problem. It is beyond the scope of this thesis to address the problem in detail, but the outlines of a solution would be that, despite appearances, there is something it is like to intend plus as opposed to quus. This is apparent when we are confronted with quus-like deviations, and intuitively respond that they are incorrect, that they do not conform to what we intended all along. Kripkenstein's mistake is to think that the phenomenology of the original intention needs to be infinitely fine-grained to establish the correctness of this intuitive preference for plus over quus – but in fact it does not. That is not how it works. Instead, the work is done by the combination of two factors: the first is a standing, implied meta-intention that, unless otherwise specifically intended, we always intend the simplest meaning that is consistent with all our other commitments; the second is the ability to intuitively perceive any quus-like deviations as being less simple than plus. Although our intuition is not infinitely fine-grained, we are able to compare plus with any of the infinitely many quus-like deviations and judge that they are less simple than plus – and therefore, given the abiding meta-intention, they cannot be what we originally intended.

the former under-determine the latter. However, Goff argues, if a powerful demon were to suddenly grant Mary perfect knowledge of the subject's phenomenal states, she would then be able to know whether the subject meant plus or quus. The facts about the subject's phenomenology would entail the facts about their mental content. Thus the idea is that the totality of the physical facts, to which Mary initially has access, do not entail the facts about meanings – but the phenomenal facts, to which she is subsequently given access, do. Therefore the phenomenal facts themselves are not contained within or entailed by the totality of the physical facts. But in the present instance, I am not assuming phenomenal intentionality – but rather arguing for it from the rule-following considerations.

But what makes it true that plus is simpler than quus? What will not work is to have a rule for determining which of two possible interpretations – plus or quus – is the simpler one. It is no good having a mathematical formula to find interpretation that minimises the conceptual deviance – that finds the line of best fit, so to speak. That would just raise the same problem: what makes it the case that *that* rule, the rule for interpreting rules, is not subject to a quus-like deviation? The only way this solution will work is if conceptual simplicity is something that we *perceive*. What makes plus simpler than quus is just that it looks simpler to us; what makes it right is that it feels right. Kripkenstein might object that whatever will seem right to me will be right. And indeed it will – but that is precisely the point; and that is how, in my view, facts about phenomenology can determine facts about meaning.

If this is right, then consciousness can determine representational content. And if it can, then the rule-following considerations provide a reason to think that it does, and therefore that strong externalism about mental content is false

Conclusion

Externalism is the thesis that the contents of our thoughts depend essentially on what the external world is actually like – that meanings ain't in the head. We can interpret this claim literally, as the thesis that contents depend on a relation to objects that are physically outside our skins – or more deeply, as a claim about the relationship between mental contents and the things they are about (wherever these are located). Moreover, we must be careful distinguish between externalism about linguistic contents – the meanings of words – and externalism about mental contents themselves. Furthermore, we must also distinguish between weak mental externalism, which claims only that some thoughts have additional, secondary contents that are externally constituted, and strong mental externalism, which claims that some thoughts only have external content. I accept linguistic externalism as something of a truism, and am not committed either way with respect to weak mental externalism. But I reject strong externalism about mental contents.

It is widely thought that certain thought-experiments, such as Twin Earth, prove that some version of strong mental externalism is true. But in this chapter, I have argued that we can use two-dimensional semantics to define truth conditions for Oscar and Toscar that are independent of their external circumstances – and therefore that the standard externalist thought-experiments do not prove the truth of externalism. This is, so to speak, a defensive argument: if successful, it leaves open the possibility that the internalist and externalist interpretations of Twin Earth are equally valid. However, I have also offered some reasons to think that externalism is false, namely that facts about mental content must be grounded in facts about consciousness. And in the next chapter, I will go further than I have in this: I will argue that two-dimensional semantics does not merely permit an internalist account of meaning, but necessitates one. The very fact that an internalist account is possible shows that externalism must be false.

Chapter 7: Ain’t Meanings In The Head? (Part II)

Introduction

In the previous chapter, I argued that Twin Earth and Arthritis do not prove the case for externalism, because it is possible to give an internalist interpretation of them using two-dimensional semantics. This seems to leave it open whether the internalist or externalist reading is correct. But in this chapter, I will argue that these considerations lead to a much stronger conclusion: the very fact that an internalist reading is even possible shows that externalism is false, and therefore the internalist account is correct. The argument is analogous to the zombie argument against physicalism. In outline: if externalism is true, then there is no possible world in which (for example) a subject on Twin Earth has Earth-like water-thoughts; but two-dimensional semantics show that it is perfectly conceivable for Toscar to have the same water thought-contents as Oscar; and, since ideal conceivability entails metaphysical possibility, it follows that this is possible; and if it is possible, then externalism is false.

In order to set out this argument, I will consider Putnam’s (1981) argument that externalism offers a way to defeat radical scepticism, because it allows me to know that I am not a brain in a vat (a ‘BIV’). I will show that Putnam’s argument does not succeed on its own terms: externalism does not defeat global scepticism; but his argument does reveal some important consequences of externalism, which will play an important role in my argument against it.

In Section 7.1, I will outline Putnam’s (1981) argument, and show why externalism, even if true, would not defeat global scepticism – it would not allow me to know that I am not a BIV. In Section 7.2, I will set out a conceivability argument against externalism that is analogous to the zombie argument against physicalism, and will show how it is possible to define 1-intensions for external-world thoughts that are the same for both a BIV and a subjectively indiscriminable non-BIV. Therefore it is necessary (and so, *a fortiori*, possible) that they have common external-world thought-contents. In Section 7.3, I will outline some of the consequences of this internalist account for our understanding of reference; in Section 7.4 I will show, contrary to externalism, that a BIV can have the true thought that it is a BIV. Finally, in Section 7.5, I will consider an objection to my argument, namely that the scenario I outline is conceivable but not really possible, and show why this objection fails.

7.1 - Could I Be A Brain In A Vat?

Many externalists have thought that externalism offers a way to defeat global scepticism about the external world. Thus, having supposedly established externalism about mental content with *Twin Earth*, Putnam subsequently (1981) argued from an externalist premise about mental content to the conclusion that I can know that I am not a brain in a vat (BIV). But is this right?

The brain in a vat thought-experiment depicts a human brain kept alive and conscious in a vat by an evil genius. Electrodes are attached to the brain, and the whole system is controlled by a supercomputer that systematically stimulates the brain, so as to induce experiences that are subjectively indiscriminable from those of a non-envatted, fully embodied brain in the real world. The result is a perfect illusion: things appear to the envatted brain exactly as they would if it were plumbed-in to the external world in the normal, embodied way.

One of the consequences of strong externalism is that, although everything is subjectively exactly the same for the BIV and non-BIV, their thought-contents relating to the external world must differ systematically. So Putnam's argument (1981) begins from an externalist theory of mental content, with the idea that a BIV and its non-BIV counterpart have systematically different conceptual schemes, with systematically different contents.

According to Putnam, we should regard the BIV’s and non-BIV’s thoughts as belonging to two distinct languages: vat-English and English, respectively. These languages are subjectively identical, but the English word ‘tree’ and the vat-English ‘tree’ have different references, and therefore subjectively indiscriminable thoughts in English and vat-English will have different truth-conditions. Therefore BIV and non-BIV can have two thoughts that are subjectively identical, but which express different propositions, and must have different contents. I have already rejected this view of content; but the aim here is to examine what would follow if externalism were true. Putnam argues that it defeats scepticism – that, if externalism is true, then I can know that I am not a brain-in-a-vat. I will argue that this does not follow, and explain why externalism, even if it were true, would not defeat scepticism.

I will consider (with some minor adjustments) one of Brueckner’s several reconstructions of Putnam’s argument, which Brueckner calls the Disjunctive Argument (Brueckner 1986, p 154).

The Disjunctive Argument is as follows:

i) Either I am a BIV (vat-speaking vat-English) or I am a non-BIV (speaking English).

[As discussed in above, this premise summarises Putnam's externalist theory of meaning. There are, according to this account, two possibilities. The first is that I am a BIV. If this is the case, then my thoughts cannot possibly be about objects in the external world – objects outside the vat – since I am not causally related to them in the right way. So what are they about? They are about whatever is in fact the systematic cause of my subjective impressions. For a BIV, the function associated with 'water' would pick out, not H_2O, and not XYZ, but whatever systematic feature of the computer program is the underlying cause of my apparent water-like experiences. Thus the BIV's water-thoughts are thoughts about some computer program analogue of water, just as Oscar's water-thoughts are thoughts about XYZ. The second possibility is that I am a non-BIV, in which case my thoughts have their normal contents. But, crucially, things seem subjectively the same to me in both scenarios. If I am a BIV, then my thoughts *seem* to express the same content as normal thoughts in English. But in this scenario I am actually not speaking English, but a different language, 'vat-English'. Vat-English perfectly resembles English, but differs systematically with respect to its contents. (And, furthermore, if I am a BIV, then I am not really *speaking* vat-English. It only seems to me that I am – I am, so to speak, vat-speaking vat-English.]

So from this theory of meaning, and considering the first disjunct, we get:

ii) If I am a BIV, then my utterances of 'I am a BIV' are true iff I am a computer program analogue of a BIV (BIV*).

[That is, if I am a BIV, then my thought that I am a BIV does not refer to my actual existence as a BIV – since I cannot refer to objects in the real world – but to the possibility that I am a computer program analogue of a BIV (which I will abbreviate to *BIV**).]

But:

iii) If I am a BIV, then I am not a BIV*.

Therefore:

iv) If I am a BIV, then my thoughts that ‘I am a BIV’ are false.

So if I am a BIV, then I cannot truly say that I am a BIV. I can only form the thought that I am a BIV*, and this thought is false.

Taking the second disjunct, we have:

v) If I am a non-BIV (speaking English), then my thoughts that ‘I am a BIV’ are true iff I am a BIV.

Therefore:

vi) If I am a non-BIV, then my thoughts that ‘I am a BIV’ are false.

So, whichever disjunct we chose:

vii) My thoughts that ‘I am a BIV’ are false.

What should we make of this? As Brueckner argues, this argument is valid – if we accept the disjunction (i), then (vii) follows. But it is important to be clear on exactly what it does show, and what it does not show.

The argument does *not* show that I could not be a BIV. The reason is that my thoughts that ‘I am a BIV’ are not univocal between the two horns of the disjunction. In the case that I am not a BIV, then my thought means what we normally take it to mean, that I am a BIV, and it is false. But in the case that I am a BIV, then my thought ‘I am a BIV’ is a thought of vat-English, not English. The English equivalent would be ‘I am a BIV* / computer program analogue of a BIV’. This thought of mine is false. But it is still true of me that I am BIV.[35]

[35] Of course, if externalism is true and I am a BIV, then this sentence will have a non-standard meaning – it will be equivalent to the claim, in English, that ‘it is still true of me that I am a BIV*’. Does this mean that, if I

Another way of putting this is that to go from (vii) to the desired conclusion:

viii) It is false that I am a BIV.

We need to invoke the disquotation principle:

My utterances of ‘I am a BIV’ are false iff I am not a BIV.

But this principle is no help since, by premise (i), my utterances of ‘I am a BIV’ have different truth-conditions depending on which scenario I am in and which language I am speaking. The disquotation principle is only available in the case where I am a non-BIV, speaking English. In the scenario where I am a BIV, speaking vat-English, then my utterance of ‘I am a BIV’ will be false iff I am not a BIV*

Thus the argument as presented here does not work: at most, it shows that if I am a BIV, then I cannot form the thought – and therefore cannot truly express the thought – that I am a BIV. But it does not follow that I cannot be a BIV. Perhaps we could adapt the argument as follows:

i) If I am a BIV, then I cannot form the thought that I am a BIV. [The externalist premise].

ii) I can form the thought that I am a BIV.

Therefore:

iii) I am not a BIV.

am a BIV speaking vat-English, then Putnam’s argument succeeds? No. This is because if externalism is true and I am a BIV speaking vat-English, then Putnam’s argument, as presented here, is itself in vat-English. In this case, Putnam’s argument (translated into standard English) is that if I am a BIV*, then I cannot form the thought that I am a BIV*. Maybe not – but it will still be true of me that I am a BIV*. The point is that, whatever meaning we assign to Putnam’s words – English or vat-English – the argument fails on its own terms.

This version of the argument is surely valid. But the problem, as Breuckner points out, is that if we assume premise (i) to be true, then we are not entitled to take premise (ii) for granted. Whilst it may seem obvious that I can form the thought that I am a BIV – am I not doing it now? – the whole point about the externalist premise is that I cannot be sure, just on the basis of access to my own subjective mental states, what their contents are. I cannot be sure that I am indeed entertaining the thought that I am a BIV, as opposed to the thought that I am a BIV*.

What the argument shows is that, if I am a BIV, then I cannot truly say that I am. But, by the same token, if I am a BIV, I cannot truly say that I am not. If I am a BIV then I cannot truly say anything at all about my situation because, according to the theory of meaning embodied in the disjunctive premise (i), I can form no coherent conception of my situation at all. But it is still true of me that I am a brain in a vat. Therefore externalism rules out the conjunction of me being a BIV and me thinking I am a BIV. But it does not rule out the possibility that I am a BIV. As Bruekner puts it:

> […] I can conclude from this that I am a normal human being rather than a BIV – and thereby lay the sceptical problem to rest – only if I can assume that I mean by 'I may be a BIV' what normal human beings mean by it. But I am entitled to that assumption only if I am entitled to assume that I am a normal human being speaking English rather than a BIV speaking vat-English. This must be *shown* by an anti-sceptical argument, not assumed in advance. [Ibid, p 160]; author's italics]

We can see this point more clearly if we consider the following three types of facts:

a) Facts about our internal physical states (i.e. brain states and so forth).

b) Facts about our subjective, internally accessible psychological states (i.e. how things seem to us from the inside).

c) Facts about mental content.

Externalism entails that our mental contents do not supervene on our internal physical states, that is, the (c) facts do not supervene on the (a) facts. But then at least one of the following must be true: either the (b) facts do not supervene on the (a) facts; or the (c) facts do not supervene on the (b) facts. But to take the first horn of the dilemma would entail that human psychology – how things are for us subjectively – does not supervene on facts about human internal physical states. Whilst this is a metaphysical possibility that some have advocated, I will not consider it further in this thesis; I assume that human consciousness supervenes on our internal physical states, the only question being whether this supervenience is metaphysical or merely causal.

The more usual response is to take the second horn of the dilemma. But the consequence of this is that two beings could be in qualitatively identical subjective psychological states and yet differ with respect to their mental content. This seems to lead to profound epistemological worries about knowledge of our own mental contents, as discussed by Boghossian (1989a), amongst others. The consequence of this is that I cannot know what my contents are just on the basis of how things seem to me subjectively. It raises the possibility that I might not know what I am thinking – that I might *think* I am thinking one thing, but in fact I am thinking something else.[36]

Therefore externalism does not solve the sceptical problem as advertised. It appears at first sight to undercut scepticism, because it entails that there is a valid inference from my mental

[36] Is this such a bad thing? It is far from obvious that we have infallible access to our own mental states. It seems highly plausible that I can be wrong about what I am thinking – and perhaps even that I can be wrong about my own phenomenal consciousness. It seems plausible that my awareness of my own phenomenal states is itself a compound state, such that I have a first-order phenomenal state, S1, and a second-order cognitive state, S2, such that S2 is the phenomenal state of it seeming to me that I am in S1. But this seems to allow the possibility that my second-order phenomenal states may not reliably track my first-order ones – and thus that it might seem to me that things seem one way, when in fact they seem another way.

However, it is one thing to accept that we can be wrong about our own mental states on occasion, and quite another to claim that there is a general sceptical problem with respect to them – that our epistemic access to our minds is no more direct or certain than our access to the external world. As Boghossian points out, externalism entails that there is a general sceptical problem with respect to mental content. I accept this point – and I take the view that in those cases where our contents are externally constituted, such as for certain linguistic contents and *secondary* mental contents, we have no more privileged epistemic access to them than we do to the external facts themselves. There is a further question about whether we even have privileged access to our own phenomenal consciousness. For example, Byrne (2015) argues for the even stronger claim that there is a general sceptical problem with respect to all mental states, including phenomenal consciousness. Since it lies well beyond the scope of this thesis to examine whether, or how, we have knowledge of our own phenomenal consciousness, I will set this issue to one side and assume that we do have such knowledge.

contents to certain corresponding facts about the external world. But the problem is that if externalism is true, then I cannot know what my mental contents are just on the basis of how things seem to me. I have no more privileged access to my own mental contents than I do to the external world itself. The fundamental problem is that although externalism entails that there is a valid inference from my mental content to certain corresponding features in the external world, it severs the link between how things seem to me subjectively and what my mental content is. So there is still no unbroken bridge from how things seem to me to how the world really is. So not only does externalism not solve scepticism about the external world – in fact, it creates a sceptical problem with respect to our own mental contents.

Although externalism – even if true – would not defeat external-world scepticism, the BIV argument does reveal an important principle: that if externalism is true, then there is no possible scenario in which a BIV has all the same primary mental contents as a subjectively identical non-BIV counterpart. This will become an important premise in the conceivability argument against externalism.

7.2 - A Conceivability Argument Against Externalism

Any version of externalism would have important metaphysical consequences if it were true. Even the linguistic and weak versions entail that some facts about content depend upon facts about the external world. But in these cases, the content affected will be merely the meanings of words (for linguistic externalism), and the additional, secondary contents of some thoughts (for weak mental externalism). Strong externalism, however, entails the stronger claim that some thoughts are such that their *only* content is constituted by a relationship to the external world. If the external world were not structured in the right way, if it did not contain the relevant objects, and if subjects were not related to them in the right way, then the thoughts in question would not be possible. They are constituted by an external relation; their content is essentially external.

Moreover, if strong externalism is true then the objects in question – and a subject's proper connection to them – are not merely *causally* necessary for certain thought-contents, but *metaphysically* necessary. According to externalism, it is not a merely contingent, empirical fact that I must be in the right relationship to certain external objects in order to possess certain mental contents: it is meant to be an essential feature of thought. The BIV scenario represents the ultimate test-case of this hypothesis: the BIV is entirely stripped of the very connections to the external world that externalists hold to be constitutive of external-world thought-contents. These connections are replaced by the vat-apparatus; real objects are replaced by simulated ones. If a BIV is capable of having all the same primary mental contents for its external world concepts as a non-BIV, then strong externalism is false. If strong externalism is not evident in the BIV case, then *a fortiori* it will not apply in the Twin Earth, Arthritis, and other less dramatic scenarios.

We can put this in terms of possible worlds: strong externalism entails that a BIV-world and a non-BIV world cannot have common external-world thought-contents – and therefore if they can have common such contents, then strong externalism is false. So the conceivability argument against strong externalism is as follows:

> i) If strong externalism is true, then there is no possible BIV-world that is a duplicate of a subjectively identical non-BIV-world with respect to primary thought-contents about the external world.[37]
>
> But:
>
> ii) For any BIV, we can define primary contents for the BIV's external-world thoughts that are the same as those of a subjectively identical non-BIV. So there is a conceivable BIV-world that is a duplicate of a subjectively identical non-BIV-world with respect to primary thought-contents about the external world.
>
> iii) Ideal conceivability entails metaphysical possibility. [This premise is necessary to pre-empt the objection that, although I can conceive of a BIV having the same primary contents as a subjectively identical non-BIV, this is not metaphysically possible and externalism is an *a posteriori* necessity.]

[37] Even if strong externalism is false, any possible BIV-world would have to differ from a non-BIV-world with respect to its *secondary* thought-contents.

Therefore:

iv) There is a possible BIV-world that is a duplicate of a subjectively identical non-BIV world with respect to internally-constituted thought-contents relating to the external world.

[In fact, if primary intensions are a function of phenomenal consciousness then *any* BIV-world *must* have the same primary contents as its subjectively identical non-BIV counterpart. But for present purposes, to make the argument against strong externalism, I only need the weaker claim that it is *possible* for BIV and non-BIV to have the same primary contents.]

Therefore:

v) Strong externalism is false.

I take it that the argument is valid, and that once premises (i) and (ii) are granted the remainder follow, and the conclusion is established.[38] I explained above why (i) is true. So the important question is whether or not premise (ii) is true. The central idea of this premise is that the relationship between a BIV and a non-BIV is analogous to the relationship between Oscar and Toscar in the *Twin Earth* case. The situation here seems more complicated, but the same principles are involved. Oscar and Toscar share a common internal content, based on their common 1-intensions, even though their external circumstances are different, and the same principles apply to BIV and non-BIV.

[38] One possible objection to this argument is that perhaps the same move could be used in the opposite direction: is there not a conceivable (and hence possible) world in which I and my BIV twin lack thoughts with the same content? And doesn't this prove externalism? But there are two problems with this objection. The first is that *if* PIT is true, then it will not even be conceivable that I and my BIV counterpart lack thoughts with the same content: the fact that we have the same subjective consciousness entails *a priori* that we have the same contents. The second problem is that even if there were a world in which I and my BIV twin lacked common contents, this would not prove externalism: externalism is not merely the claim that there is a world in which we lack common contents – it entails that there is no world in which we do have common contents.

Suppose that a BIV and a non-BIV both have the thought ‘There is a tree’, when confronted with subjectively identical tree-impressions in appropriate circumstances. Internalism implies that there is a common content to these thoughts – that is, that there is a sense in which they both have the same truth-conditions. How does this work?

At first sight, it might seem that the non-BIV’s thought in these circumstances is true, and BIV’s thought is straightforwardly false. The non-BIV is confronted by a tree, whereas the BIV is confronted by a mere simulation of a tree. This would imply that both thoughts do indeed have the same truth-conditions, requiring the presence of a tree, and not a simulated tree. So far so good for internalism.

But this is too quick. We must be careful to assess the truth or falsity of the BIV’s thought in terms of the BIV’s own tree-concept, not that of its non-envatted counterpart. So on second sight, we might take the view that their respective thoughts will be made true by the presence of whatever is the systematic cause of their tree-like impressions. For a non-BIV, this is a tree; for a BIV, this is a computer simulation of a tree (that is, a *tree**). So now the two thoughts seem to have different truth-conditions – to be about different things – and therefore to have different contents. This appears to be a problem for internalism.

But this is not the end of the story. In fact, an internalist analysis can allow the possibility that the BIV’s thought is made true by the presence of simulated tress, and the non-BIV’s thoughts by the presence of real trees. Recall that, in the *Twin Earth* case, Oscar and Toscar’s respective thoughts ‘there is water’ can both have the same content, even though they are made true by the presence H_2O and XYZ, respectively. In that case, the solution was that the concept ‘water’ is a comparative function that compares the watery substance presently in front of the subject to the watery substance the subject is habituated to, and declares them to be the same. Thus the thought ‘there is water’ is true just if the watery substance in the present circumstance of evaluation is indeed the same as the substance that the subject is habituated to, whether that be XYZ or H_2O

A similar analysis is possible in the BIV case. If the primary content of the tree-concept is a function that picks out whatever is the systematic, underlying cause of my tree-like impressions, then this function will be satisfied by trees* if I am a BIV, and by actual trees if I am a non-BIV. In either case, my thought ‘there is a tree’ will be true whenever I am confronted by whatever is in fact the systematic underlying cause of my tree-like impressions, whether that be tree or tree*. So for BIV and non-BIV alike we can define a common, internally constituted function that yields the same truth conditions for both their respective tree-thoughts. It is just that BIV and non-BIV, like Oscar and Toscar, have different histories of causal interaction with the external world; so they need to be exposed to different items for their contents to be satisfied, and their thoughts to be true.

If this analysis is right then the thought ‘there is a tree’ has something like the following structure:

> ∃x ∃y [(Is the type of the systematic cause of my tree-like impressions) x & (Is the particular object I am now indicating) y & (Is a particular object of the type) y, x]

This shows that it is possible to give an internalist account of external-world thought-contents that will yield the same primary content for BIV and non-BIV. But of course the same primary content will yield different secondary contents for BIV and non-BIV, given their different causal histories. For the BIV, the secondary content of this function will be a tree*; for the non-BIV, it will be a tree.

This analysis – essentially that BIV is to non-BIV as Oscar is to Toscar – shows that premise (ii) is true, and therefore the argument against strong externalism will succeed. It is *possible* for BIV and non-BIV to have the same external-world thought-contents – and this is sufficient to refute strong externalism.

7.3 – Internalism and Reference

This analysis provides a plausible solution to a deep problem for internalist accounts of meaning, which I alluded to in Chapter 6, Section 6.2. This concerns how any internally constituted meaning can refer to objects in the external world at all. The problem arises if we suppose that some form of internalism is true, and that the basic meanings of external-world concepts are primary intensions fixed by something like descriptions – so an object, x, is F if and only if it is both G and H. Then what fixes the content of the predicates G and H? We will have to define them by description in terms of further predicates I, J and so on. But this threatens to become either an infinite regress that never hits bottom, or a closed circle of meanings defined in terms of each other. How, on this model, do our concepts manage to refer to anything in the external world at all? The analysis I have proposed here solves this problem by defining some of our external world concepts as referring to whatever is the systematic, external cause of certain experiences – so an object, x, counts as a tree just if it is of the type that is the systematic, external cause of our tree-like experiences, and so forth.[39]

This analysis does raise further issues, which would take me beyond the scope of this thesis, but which are worth mentioning. The first is that it employs the concept of a cause in basic external-world concepts. But where do we get the concept of a cause from? There are a couple of possibilities: one is that it is an *a priori* concept (in the Kantian sense), being a necessary condition for the possibility of any structured, representative experience; another is that causation is built-in to our perceptual experience of the world, and that we take it from our experience and then apply it to the relationship between that experience and whatever is its external cause (Kant would not approve of this).

A second, more serious issue concerns how we refer to our own experiences. For the sake of simplicity, I have defined the concept *tree* as referring to whatever is the systematic, external cause of my tree-like experiences. But what is a tree-like experience? It cannot just be an experience of the type caused by trees – that is obviously circular; nor does it help to define it in terms of any external objects – that will just lead us back to the closed circle of descriptions. For this strategy to work, therefore, we need to be able to refer directly to our experiences: a tree-like experience is an experience like *this*.

But how we refer to our own experiences is a controversial matter; it is not universally accepted that we can refer to them directly. Byrne (2019) argues that how things seem to us in experience

[39] The alternative is that descriptive contents are ultimately grounded in primitive concepts that refer directly. But this raises the question of what type of objects these supposed primitive concepts refer to. There seem to be two possibilities: either they refer to external objects, or to our own subjective states. If they refer to external objects, then the resulting account of meaning will be incompatible with internalism, for we just smuggled externalism back in to the picture; if they refer to subjective states, then our external-world concepts will be structural descriptions of our own phenomenology, and phenomenalism or subjective idealism looms. The challenge is to find an account of how our concepts latch on to the world that respects both a thorough internalism about mental content and a commitment to realism about external objects; my view aims to encompass both.

is characterised by how the world, objectively, is presented to be. If pressed to describe how things seem to us – the phenomenal character of our experience – we describe how the world seems to be. Byrne's aim is to argue that perceptual experience (veridical or not) requires the existence of external objects. Evidently, I do not agree with this conclusion. But it is highly plausible that perceptual experience does require the appearance of external objects – that phenomenal character is inextricably linked to the structured presentation of objects and properties. But then how can we describe our experiences without using ordinary external-world concepts? But perhaps this is asking too much of an internalist account. In order to allow the internalist analysis of external object concepts, we do not need to describe the content of our experience in object-neutral terms – we merely need to refer to it in object-neutral terms. Perhaps we can refer directly to *this* experience without having to describe what this experience is like.

A related issue is that, even if we are able to refer directly to our own experiences, it is highly plausible that reference to objects precedes reference to experiences. Infants do not first learn to refer to their own phenomenal states, and only subsequently infer the existence of external objects as their cause. On the contrary, they first learn to refer to objects as they are presented in experience, and only much later acquire the ability to refer to their experiences as such. This suggests that there is a sort of genealogy of reference at work: first, we acquire the ability to refer to objects as they appear to us in experience; then we learn to refer to our experiences as such; finally, and only then, can we refer to objects as they are in themselves, as the external causes of our experience.[40]

7.4 – What is a BIV Thinking When it Thinks it is a BIV?

The internalist analysis of external world concepts shows how BIV and non-BIV can have the same thought-contents; this allows the argument against externalism to be made. But there does

[40] This suggests that infants and naïve realists, who do not posses object concepts of the form 'the systematic external cause of my experience…' cannot refer to external objects as such, but only to the appearance of the object in experience. In practice this makes no difference because our experiences are systematically connected to objective reality, and everything seems to naïve realists as if they were referring to external objects themselves.

seem to be a worry that the internalist analysis goes too far, and entails the counter-intuitive consequence that a BIV's external world beliefs are generally true. Just as Toscar's water-related beliefs will be true in virtue of the distribution of XYZ, so a BIV's tree-related beliefs will be true in virtue of the BIV's relations to the systematic, underlying causes of its tree-like impressions – namely trees*. But surely the BIV is deceived? Surely its beliefs about the external world are systematically, generally false?

Suppose that a BIV, like Neo in *The Matrix* (1999, Warner Bros, dir. The Wachowski Brothers), is released from the simulation and awakened into the real world. What should it make of its new surroundings? And what should it now make of the simulated world, from which it has just awakened? Surely it will – and should – judge that its present surroundings are the real ones, and its previous, envatted experience was illusory. Hence, when confronted in the real, meta-world by actual trees (as opposed to mere trees*) it will (and should) think: 'When I was in the Matrix, I thought I encountered trees, but I now see that they were merely simulations of trees. *These things, in the real world* are the real trees.'

So which is right – the analysis that renders a BIV's external world beliefs generally true (in the appropriate circumstances), or the intuition that says they are globally, collectively false? What does seem harmless enough is to regard the BIV's beliefs as true relative to the BIV's simulated reality. A BIV's external world beliefs are true-in-the-Matrix in much the same way that things can be true in the context of a computer game, say, that are not true in the real world. For example, it can be true within the context of a computer game that a particular character is in a particular environment, even if it is not true outside of this context – because the character and environment do not exist in the real world. (They do exist in the real world as simulated or fictional objects, created by developers, but they do not exist as flesh-and-blood physical objects). But it is one thing to grant the BIV's beliefs a relative degree of truth, valid only within the scope of a sort of fictional discourse – it is quite another to allow that the BIV's external world beliefs are true *simpliciter,* that it has access to reality as it really is.

In 'The Matrix as Metaphysics' (republished in Chalmers 2010, pp 455-494), Chalmers argues for the latter, stronger claim: that a BIV (generally) has true external-world beliefs, and that it has access to the real world. Of course, the BIV does not have access to the meta-reality that supports both its own existence and its experiences. It does not have epistemological access to the vat, or the machines controlling the Matrix, and so forth – but, argues Chalmers, this is not what is required for it to have access to reality and for its external-world beliefs to be generally true.

The central idea of Chalmers's argument is that scenarios such as the Matrix, brains-in-vats and so forth are not really sceptical hypotheses but metaphysical ones. If it turns out that I am a BIV or living in the Matrix, then my everyday beliefs will be generally true and made true by my appropriate relation to the systematic, external causes of my experiences. But it will be a metaphysical truth that the systematic, external causes of my experience are elements of a computational program controlled by evil scientists or machines, as the case may be. Chalmers calls this the *Metaphysical Hypothesis,* and it consists of three sub-hypotheses: the Computational Hypothesis, which says that 'microphysical processes throughout space-time are constituted by underlying computational processes' (p 460); the Creation Hypothesis, which says that 'physical space-time and its contents were created by beings outside physical space-time' (p 461); and the Mind-Body Hypothesis, which says that 'my mind is […] constituted by processes outside physical space-time and receives its perceptual inputs from and sends its outputs to processes in physical space-time' (p 462). Of all these hypotheses, and of the Metaphysical Hypothesis as a whole, Chalmers claims that each is coherent, cannot be ruled out by what he knows, and – crucially – is not a sceptical hypothesis.

But there is a problem with this argument, in my view, which is that Chalmers is too quick to declare that the Creation Hypothesis is not a sceptical hypothesis. I think the question of whether the Creation Hypothesis is sceptical or metaphysical very much depends on who or what the creator is. If the creator is God, so that the physical process in space-time – to which our minds are related in experience – are constituted by computational processes grounded in a creator God, then it will be a metaphysical truth that reality is a sort of Matrix in the mind of God. There would be no sense at all in which we are the victims of a global deception. True, we would not have direct epistemological access to things as they are in themselves; we would experience reality as mediated by our cognition. But the important point is that the computational process, which constitute the inputs and outputs connecting our minds to the outside world, represent the base level of (created) reality. And I think there is an underlying, implied assumption in all of our external-world beliefs that the systematic external causes of our experiences are part of the base level of reality. If, on the other hand, the creators of (our) space-time are machines or evil scientists, who exist in their own space time and are subject to their own laws of physics (which may or may not resemble our own), then we do not have even indirect access to the base level of reality. The computational processes that underly our physics and our space-time will be one level removed from the computational processes (or whatever else) constitutes the physics and the space-time which the machines inhabit. In this case, the implied assumption in all of our external-world beliefs will be false, and therefore our beliefs themselves will be systematically, globally false.

The same question arises if we consider the BIV's thought that it is a BIV. Just as we might intuitively think that there is a sense in which a BIV's tree-related thoughts are false, so we might think that a BIV can think the *true* thought that it is a BIV. But how can a this be the case? Surely its thought 'I am a BIV' will be true just if it is an object of the type that is in fact the systematic, underlying cause of its BIV-impressions – namely a BIV*. But if it is a BIV, then it is not a BIV*, and its thought 'I am a BIV' will be false.

What is the solution? Let us consider a revised version of the BIV thought-experiment, to see what a BIV is really thinking when it thinks it is a BIV. We can imagine an envatted brain that has been vat-reading vat-Putnam, and vat-watching *The Matrix.* The mad scientist controlling the BIV then arranges for it to be presented with experience of an apparent BIV (after vat-watching The Matrix, the BIV vat-wanders into the research department of the local hospital*, and encounters what appears to be row after row of living brains in vats, all connected by wires and electrodes to a controlling supercomputer). So the BIV begins to question whether or not it is itself in fact a BIV, and ultimately comes to believe that it is. Thus it entertains a thought with the apparent content 'I am a BIV'. The question is: what does this thought mean? What is its real content? Is it thinking that it is a BIV? Or that it is a BIV*? In the first case, its thought will be true; in the second, it will be false.

To shed some light on this, we continue the thought experiment in the following way: the scientist decides to lift the envatted consciousness from its simulated world and reveal to it the truth of its situation. So she puts the brain into a coma, removes it from the vat, and transplants it into the living (but hitherto brainless) human body kept to one side for this very purpose. The scientist wakes the brain, which is now fully plumbed-in to its new body, and has working eyes, ears, speech, limbs and so forth. The now-embodied brain is no doubt confused by its new situation, but the scientist explains what has happened, pointing to the vat and the computer to which the brain was formerly connected.[41] The now-embodied brain is able to walk around and interact with the apparatus. There is a period of adjustment as it becomes immersed in its new environment. Eventually it is convinced about what has happened.

As it reflects on its experiences, what judgement would it make regarding its former, envatted situation? Does it now judge that its previous, envatted thoughts that it might be a 'BIV' were in fact true – or were they false? Does it now think: *when I was a BIV, I thought I was a BIV* - which of course I wasn't?* Or does it think: *when I was a BIV, I thought I was a BIV – and it turns out I was right all along?*

This depends on how the formerly envatted brain understands the content of its 'BIV' concept. There are two different ways to interpret the BIV's primary intensions. According to the first interpretation, the BIV's thought will be false; according to the second interpretation, it will be true. So we have the following two ways to interpret the BIV's 1-intensions:

<u>Interpretation 1</u>: If by 'BIV' it meant something like 'whatever is the systematic external cause of my BIV-representations', then the newly embodied brain would have learnt that the reference of its 'BIV' concept is in fact a BIV*. And, since it was not in fact a BIV*, it would have to judge that its envatted thoughts were false. This is analogous to interpretation of *Twin Earth* where Oscar's 'water' concept means (roughly) 'whatever microphysical substance instantiates the watery properties in my normal environment'. If that is indeed Oscar's 1-intension for 'water', then, on being told that he is in fact on Twin Earth and the substance in front of him is in fact XYZ, he would concede that it is not in fact 'water' as he understands the term. It is important to emphasise that, even if this is how the newly embodied brain were to interpret the situation, it does not mean that externalism is true. We can still allow for a sense in which an envatted and a non-envatted brain can both have the same 1-intensions, and therefore the same primary content, for their BIV-concepts. On this interpretation, for an

[41] Of course, if externalism is true, then the evil scientist cannot even hold a meaningful conversation with the newly-embodied brain, having released it from the vat. The scientist, after all, speaks English – but the brain only knows vat-English. It will seem to both parties as if they are having a meaningful conversation (since English and vat-English sound exactly the same), but their words will have completely different contents, and they will just be talking past one another.

individual to think ‘I am a BIV’ is to for it to judge ‘I am the same sort of thing as the systematic external cause of my BIV-representations’. This thought has the same 1-content whether the thinker is envatted or not – it just means that *whatever* the external cause of my BIV-representation, I am *that* sort of thing (and of course, this thought will always be false, whether I am a BIV or not). The 2-content, however, will vary depending on whether or not the thinker is envatted at the time.

So on this interpretation, the 1-intension of the thought ‘I am a BIV’ has something like the following structure (as per the tree example, above):

> ∃x ∃y [(Is the type of the systematic underlying cause of my BIV-impressions) x & (I am identical to the particular object) y & (Is a particular object of the type) y, x]

The problem with this interpretation is that this is clearly not what I mean when I entertain the thought that I am a BIV – and nor is it what I would mean if I were indeed a BIV. When, in this scenario, I have an impression of a BIV and think ‘maybe I am a BIV’, what I am *not* thinking is that maybe I am the same sort of thing as whatever is causing my present BIV-impression. That is to entirely miss the point of my speculation. The possibility I am entertaining is that *maybe something like that is going on with me*. This is captured in the second interpretation.

<u>Interpretation 2</u>: The newly embodied brain judges that it had been right all along. But then its ‘BIV’ concept cannot have the content ‘whatever is the systematic external cause of my BIV representations’. What is the content then? In this scenario, the function of a subject’s ‘BIV’ concept is not to compare the subject to the external cause of its BIV-representations, and judge that they are the same type of thing, but rather to compare the subject’s situation with respect to its experiences to what appears to be going on what it encounters an apparent BIV. When, in its envatted state, the brain encountered what appeared to be a BIV (and which was in fact a BIV*), and thought ‘I am a ‘BIV’’, it is not thinking (as per interpretation 1) ‘I am the same

sort of thing as the cause of this BIV-experience'. Rather, it is thinking 'What *seems* to be going on with this 'BIV' is *really* going on with me'.[42] And this thought would be true.

This is trickier to formalise, but on this interpretation, the 1-intension of the thought 'I am a BIV' has something like the following structure:

In simple English, we have:

a) I exist and I have experiences.

And

b) My experiences are such that apparently: there exists a BIV, which has experiences.[43]

And

c) The actual relationship between me and my experiences is the same as the apparent relationship between the apparent BIV and its apparent experiences.

And formally we have:

∃x ∃y [(Are my experiences)x & (I am identical to)y]

&

[42] The thought is that I stand in the same relationship to my experiences and representations as the apparent BIV stands to its (apparent) experiences and representations. So on this interpretation, my 'BIV' concept is not just a two-place function – it is in fact a four-place function.

[43] NB it is *not* part of my thought that the BIV I encounter in experience really exists, because in thinking that I am really a BIV, I must also think that the apparent BIV is not itself real.

> I have experiences such that apparently: ∃z ∃w [(Is a BIV) z & (Are the experiences of)w, z]
>
> &
>
> (x : y) = [the appearance of: (w : z)]

The second interpretation seems to me a better account of what we are actually thinking when we entertain the thought that perhaps we are brains-in-vats. In the film The Matrix, when lead character Neo is released from the Matrix, he realises that Morpheus has told him the truth; he does not accuse Morpheus of having deceived him (‘You told me I was in The Matrix*, but it turns out I was in The Matrix!’).

However, it is important to emphasise that the argument against externalism stands whichever interpretation is right. Equally, it doesn’t matter whether Chalmers is correct, and The Matrix is a metaphysical hypothesis, or I am right, and it is a sceptical scenario – in either case, we can show that it is possible for a BIV to form internally-constituted external-world thoughts. And if it is possible, then strong externalism is not true of our external world concepts in a BIV scenario. But if strong externalism is not true in this case, then it is not true of any mental content. Therefore we should conclude that strong externalism is false.

7.5 - An Objection: BIVs Do Not Have Concepts At All

In this section, I will consider an objection to the argument I have set out in the preceding sections, and explain why the objection does not work. This objection is analogous to that made against the zombie argument by *a posteriori* physicalism: that the scenario in question may be primarily conceivable, but it does not represent a genuine metaphysical possibility. In the present case, the objection is that the scenario I have described, in which a BIV has external-world thought-contents (whether about trees or BIVs or whatever) that are common with those of a subjectively indiscriminable non-BIV, is 1-conceivable but not 2-possible.

The idea is that externalist considerations do not just affect the contents of our concepts, but what it actually means to possess a concept at all. According to this view, our concept 'concept' functions somewhat like our concept 'water'. We have an ordinary, surface notion of what a concept is – just as we have a surface notion of what water is. Our surface notion of a concept is (roughly) that it is something mental, representational, and so forth – just as our surface concept of water is that it is a watery substance. But – and this is the crucial point – just as our concept 'water' rigidly designates the microphysical substance that actually occupies the relevant watery role, so our concept 'concept' rigidly designates those structures in the world that actually play the relevant role in mental representation.

But what are these actual world structures that play the relevant role in mental representation? One possibility is that they are brain states or processes. But this is no help to the externalist cause, since BIV and non-BIV will have the same brain states. However, there is an alternative possibility, which is that the relevant structures are not merely internal physical states of the subject, but physical states that are appropriately connected to the external world via relevant causal-historical connections. If this is right, then it opens up the possibility that a BIV will not merely fail to have the *same* concepts as a non-BIV, but it will not have concepts *at all* in the sense that a non-BIV has them. For a non-BIV, to have a mental representation is to be in an appropriate internal physical state that is causally connected to the external world in the relevant ways. If this is indeed what concepts are in the non-BIV world, then a BIV will not have concepts at all in this sense – and, *a fortiori*, it cannot have the same concepts. We might well say that, rather than concepts, a BIV has twin-concepts, in the same way that Twin-Earth has twin-water.

We can put the argument as follows:

> i) The concept 'concept' functions much like the concept 'water'. Just as 'water' rigidly designates whatever microphysical substance actually instantiates the watery properties, so the term 'concept' has a primary sense that refers to mental representation in general, and a secondary sense that refers to whatever physical structures in the world actually play the relevant roles in mental representation.
>
> ii) Whilst it is perfectly possible that BIV and non-BIV have the same 1-concepts (that is, they have the same concepts in the primary-intension meaning of the term 'concept'), this is not enough for the argument against externalism to succeed. For that argument to work, BIV and non-BIV must have the same 2-concepts – that is, they

must have in common whatever physical structures play the relevant role in mental representation in the actual world.

iii) In the actual (non-BIV) world, the physical structures that play the relevant role in mental representation are not merely internal physical structures (such as brain states), but internal structures that are connected to the external world by appropriate causal- historical relations.

iv) Since a BIV does not have the appropriate causal-historical connections to objects in the external world, it cannot have 2-concepts at all.

v) *A fortiori*, a BIV cannot have the same 2-concepts as a non-BIV.

How to respond to this objection? I will concede, for the sake of argument, that the concept 'concept' has a 2-intension that picks out whatever physical structures play the relevant role in mental representation in the actual world. (In fact, I think this is quite plausible). I will also concede that, in the actual world, the structures which play the relevant role in mental representation are not just internal structures such as brain states, but include an appropriate causal relationship to objects in the external world. (In fact, I think it is more plausible that they are just brain states, but it is not important for present purposes, so I will concede the point). So it follows from this that a BIV will not have 2-concepts at all – any more than Twin-Earth has 2-water.

So this objection may show that there is a sense of the term 'concept' in which the BIV does not have concepts at all, let alone the same concepts as his non-BIV counterpart. But – and this is the crucial point – this is not the relevant sense of the term 'concept'. It is not the sense that matters as far as the argument against externalism is concerned. We may use the term 'concept' to refer to such structures, but that is not its only legitimate, or even its most natural sense.[44] We also use it to refer to a primary intension *per se*. Indeed, the primary intension of the concept 'concept' is a primary intension. And the argument against externalism rests on the idea that BIV and non-BIV have the same 1-intensions for their external-world thoughts. But what is a

[44] And hence it is wrong to assert that the term 'concept' *rigidly* designates a particular type of structure in the actual world.

1-intension? It is a function of how things seem from the point of view of the subject, of their subjective consciousness. Hence 1-intensions are like phenomenological properties – whatever they seem to be to the subject, that is what they are. And, just as there is no difference between the 1-intension and 2-intension of our phenomenological concepts, so there is no difference between the 1-intension and 2-intension of the concept '1-intension'. There is no possibility that the concept '1-intension' refers to one type of physical structure in a non-BIV world, and another type in the BIV world. Hence there is no possibility that the non-BIV world contains 1-intensions, but the BIV world does not. If they are subjectively indiscriminable then, by definition, they contain the same 1-intensions, which is all that is required for the argument against externalism.

We can see all this illustrated clearly in the case of *Swampman* (Davidson, 1987). In Davidson's thought-experiment, he is walking in a swamp and is killed by a bolt of lightning. At the same time, nearby in the swamp, another bolt of lightning spontaneously causes adjacent matter to be arranged in a molecule-for-molecule perfect replica of the now deceased Davidson, and which Davidson (in the real world) called 'Swampman'. So here is the question: does Swampman, at the instant of his inception, have the same mental contents that Davidson had at the moment of his death? We assume that Swampman is not a zombie. He has the exact same microphysical constitution as dead-Davidson, and we assume that consciousness supervenes on microphysics. Of course, being a dualist, I will insist that consciousness supervenes causally on the microphysics, and not metaphysically. But that is not the issue at present; the important point is that Swampman and dead-Davidson are subjectively indiscriminable. So what follows?

According to Davidson, Swampman does not have the same mental contents as dead-Davidson, for reasons similar to those set out above. Swampman does not possess the same concepts as dead-Davidson, because part of what constitutes having a representation is being in the right causal relationship to objects in the world; and Swampman does not have the right relatedness to the world. Of course, he does have the same consciousness as dead-Davidson. So he will go about the rest of his day as Davidson, had he lived, would have done. And it will seem to everyone with whom Swampman interacts – and indeed, to Swampman himself – that he posses the same mental contents as dead-Davidson. But, according to Davidson, this appearance is misleading. He may appear to recognise objects, people etc from a time before Davidson went into the swamp, but he cannot actually do so, because he (Swampman) never really cognised the objects in the first place. Perhaps after a period of time Swampman will indeed acquire concepts, but only once the relevant causal historical connections between his internal physical states and the external world have been established for the former to count as representations of the latter. In this respect, he may be like Oscar undergoing a slow-switching of his water-concept after a period of time on Twin Earth.

Is Davidson right? It is plausible that he is right about Swampman's 2-concepts. (It is important to be clear that it is not the secondary intensions of Swampman's concepts that is the issue here, but rather the secondary intension of *our* concept 'concept', and what features of Swampman are required to instantiate it.) So plausibly, the 2-intension of our term 'concept' refers to the sort of externally connected functional states that were present in dead-Davidson (when he was alive), but are not present in Swampman (at least not at first). If that is right, then Swampman may not have any 2-concepts at all, let alone having the exact same ones as dead-Davidson.

But – and this is the crucial point – even if Davidson is right about the secondary intension, it remains the case that the primary intension of the concept 'concept' is just a primary intension. Even though (if Davidson is right) our 2-concepts are constituted by external facts, our 1-concepts are not. And these are the same for Swampman as they are for dead-Davidson, because both have the same consciousness. One way of looking at this is that both Swampman and the BIV are in the opposite situation to philosophical zombies. Zombies replicate the physical states and functions (both internal and external) of normal people, and therefore arguably have the same 2-representations (i.e. the same functional representations); but they lack consciousness, and therefore do not have 1-representations (i.e. conscious representations) at all. Swampman and BIVs, on the other hand, fail to replicate the external physical relations of normal humans (although they do replicate the internal physical conditions). So it is at least plausible that they do not have the same 2-representations as ordinary people. But, because they have the same consciousness, they do have the same 1-representations.

So this objection, whilst it makes a plausible case about the secondary intension of the concept 'concept', does not threaten the argument against externalism.

Conclusion

It is sometimes argued that externalism, if true, would undercut global scepticism about the external world. But I have argued that this argument does not work. At its heart is the idea that a thinking subject who is the victim of a global deception about the external world – such as a brain-in-a-vat – would have systematically different thought-contents from a subjectively identical thinker who was not deceived (that is, who is correctly connected to the external world). So if externalism is true, then a BIV cannot form the true thought that it is a BIV. Unfortunately, this is no help to me when I wonder whether or not I am a BIV. The problem is that if externalism is true, then I cannot know what my own thought-contents are. I do not know whether, being a non-BIV, I am thinking falsely that I am a BIV – or, being a BIV, I am thinking falsely that I am a BIV*. Either way, my thought will be false. But in one case, I will be a BIV.

The moral of this is that whether externalism or internalism is true, there is no valid inference from how things seem subjectively to how they are objectively. Internalism allows an inference from how things seem subjectively to mental content, but denies an inference from mental content to the objective facts; externalism allows an inference from mental content to the external facts, but denies an inference from how things are subjectively to mental contents. Unfortunately, externalism does not offer a way to solve the sceptical problem.

However, there is an important conclusion to be drawn from the BIV scenario. This is that strong externalism does indeed rule out the possibility of a BIV having the same internally-constituted external-world thought-contents as a non-BIV

So if it turns out that a BIV can indeed think the same thoughts as a non-BIV, this will show that the externalist thesis is false. And this is indeed what I have argued in this chapter: that it is possible for BIV and non-BIV to have the same external-world primary thought contents. In fact, it is *necessary* that a subjectively indiscriminable BIV and non-BIV will have common external-world thought-contents. This is because they will have the same 1-intensions, and 1-intensions determine truth conditions that are independent of the identity of their references. *A fortiori*, the very thing that strong externalism rules out is in fact possible. Therefore we should conclude that strong externalism about mental content is false – and that meanings really are in the head.

Conclusions

The central claim of this thesis – the idea on which everything else rests – is that there are no strong *a posteriori* necessities. I have argued that there are no strong necessities because they are impossible, and they are impossible because they are ultimately incoherent. It is merely *prima facie* conceivable that there could be an ideally conceivable scenario that is not metaphysically possible – but there is ultimately no coherent way to formulate this apparent hypothesis. Therefore modal rationalism is true, and ideal conceivability entails metaphysical possibility.

This has important consequences for the metaphysics of mind. In particular, modal rationalism is an essential premise in the two-dimensional argument against physicalism. I have argued that this argument is successful – that it shows that there is a possible world that is a minimal physical duplicate of the actual world, but which is devoid of consciousness, and therefore that physicalism is false. So if P represents the totality of microphysical facts, and Q represents the phenomenal facts, and T represents a 'that's all' clause, and an arrow represents metaphysical entailment, then we can represent this conclusion as follows:

1) PT =/=> Q

For the purposes of this thesis, I have discounted panpsychism and its variants. This means that if physicalism is false, then some form of dualism is true. The minimal form of dualism is property dualism. Property dualism is consistent with the existence of causal laws that connect microphysical facts and consciousness. Whether or not there are such laws is a question of empirical fact; but I take it that it is highly likely that there are. In that case, it is plausible that consciousness supervenes causally, although not metaphysically, on fundamental physics. If L represents the causal laws that connect consciousness and fundamental physics, then we have the following:

2) PT + L ➔Q[45]

[45] I take a non-Humean view of the laws of nature. So there is a possible world that consists of just L on its own, with no P (and no Q).

Modal rationalism also leads us to the conclusion that strong externalism about mental content is false. Moreover, I have argued that it is false for the same reason as physicalism: both theories entail false claims about the space of possible worlds. Specifically, strong externalism entails that a BIV world will not have the same external-world thought-contents as a non-BIV world. But, using the two-dimensional semantic framework, it is possible to define internally constituted thought-contents that are the same for both BIV and non-BIV worlds. Since modal rationalism is true, it follows that there are corresponding possible worlds, and therefore that strong externalism is false. The argument against strong externalism, however, does allow thoughts to have additional, secondary contents that are externally constituted and which differ between BIV and non-BIV worlds.

If internalism is true, then it seems natural to think that mental contents are determined by phenomenal consciousness. Although it is beyond the scope of this thesis to explore the phenomenal intentionality theory in great detail, I have offered several reasons to think some version of it is true – most notably, that it is the only way to address certain concerns about meaning and rule-following. If we accept some form of phenomenal intentionality, and C represents the totality of facts about primary mental contents, then we have the following:

3) $Q \rightarrow C$

These, then, are the three main conclusions of this thesis: first, that modal rationalism is true; second, that the two-dimensional argument shows that physicalism is false; third, that an analogous two-dimensional argument shows that strong externalism is false. I believe that these conclusions also shed light on the nature of the ordinary macroscopic world that we encounter in experience, and how it relates to both consciousness and fundamental physics. So I will end this thesis by outlining a somewhat speculative view of these matters.

If M represents the totality of macroscopic facts that are not conscious, then the macroscopic world as a whole (including consciousness) is represented by [M + Q]. But what is the relationship between M and P? I have argued that there are two possibilities. The first is that P *a priori* entails M, in which case we simply have:

4a) $P \rightarrow M$

But the second possibility is that some form of what I have called semi-antirealism is true, and M itself contains elements that are essentially related to consciousness (even though we have already defined M so that it does not contain any elements that are themselves conscious). In this case, P on its own will not be metaphysically sufficient for M. However, it is plausible that the following will hold:

4b) $P + C \Rightarrow M$

C will contain as a subset the translation manual, V, that maps P-facts on to M-facts. It is also important to note that in conditional (4b), C represents the mental contents of the world's inhabitants. This means that the output M is only entailed relative to the cognition of these inhabitants – hence why I have described this view as semi-antirealist. A God-like observer of the world need not have a cognitive scheme represented by C. But such an ideal conceiver, knowing that the inhabitants of the world have representational content C, (which contains translation manual V) could deduce *a priori* that the macrophysical facts are M *from the point of view of the world's inhabitants*.

Now if we combine the conditionals (1), (2), (3) and (as the case me be) either (4a) or (4b), then we get the following:

5) $PT + L \Rightarrow M + Q$

On the right hand side is the everyday world of macroscopic objects and consciousness; and on the left hand side is the realm of fundamental physics and the hypothesised psycho-physical connecting laws. So this conditional represents the metaphysical relationship between the everyday world and the fundamental realm on which it supervenes.

Bibliography

Bach, Kent (1981). ‘What's in a name?’ Australasian Journal of Philosophy 59 (4):371 – 386.

Balog, Katalin (1999). ‘Conceivability, possibility, and the mind-body problem.’ Philosophical Review 108 (4):497-528.

Berto, Francesco, "Impossible Worlds", The Stanford Encyclopedia of Philosophy (Winter 2013 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/win2013/entries/impossible-worlds/>.

Bennett, Karen (2008). Exclusion again. In Jakob Hohwy & Jesper Kallestrup (eds.), *Being Reduced: New Essays on Reduction, Explanation, and Causation*. Oxford University Press. pp. 280--307.

Blackburn, Simon (1984), Spreading the Word. Clarendon Press.

Block, Ned & Stalnaker, Robert (1999). ‘Conceptual analysis, dualism, and the explanatory gap’. *Philosophical Review* 108 (1):1-46.

Boghossian, Paul (1989a), “Content and Self-Knowledge,” *Philosophical Topics,* 17: 5–26.

Bourget, David and Mendelovici, Angela, "Phenomenal Intentionality", *The Stanford Encyclopedia of Philosophy* (Fall 2019 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/fall2019/entries/phenomenal-intentionality/>.

Brueckner, Tony, "Skepticism and Content Externalism", *The Stanford Encyclopedia of Philosophy* (Spring 2012 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/spr2012/entries/skepticism-content-externalism/>.

Brueckner, Anthony L. (1986), “Brains in a vat”. *Journal of Philosophy* 83 (3):148-167.

Burge, Tyler (1979), “Individualism and the Mental,” in French, Uehling, and Wettstein (eds.) *Midwest Studies in Philosophy*, IV, Minneapolis: University of Minnesota Press, pp. 73–121.

Byrne, Alex (2015). “Skepticism about the Internal World,” *The Norton Introduction to Philosophy,* ed. G. Rosen et al., W. W. Norton.

Byrne, Alex (2019), “Perception and Ordinary Objects,” *The Nature of Ordinary Objects,* ed. J. Cumpa and B. Brewer, Oxford.

Chalmers, David J. (1996). ‘The Conscious Mind: In Search of a Fundamental Theory.’ New York: Oxford University Press.

Chalmers, David J. & Jackson, Frank (2001). ‘Conceptual analysis and reductive explanation’. *Philosophical Review* 110 (3):315-61.

Chalmers, David J. (2010). ‘The Character of Consciousness.’ New York: Oxford University Press.

Chalmers, David (2012). ‘Constructing the World.’ Oxford University Press.

Chalmers, David J. (2014). ‘Strong necessities and the mind–body problem: a reply.’ Philosophical Studies 167 (3):785-800.

Choi, Sungho and Fara, Michael, "Dispositions", *The Stanford Encyclopedia of Philosophy* (Fall 2018 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/fall2018/entries/dispositions/>.

Coleman, S., & Alter, T. (2018). Panpsychism and Russellian Monism. In W. Seager (Ed.), *Routledge Handbook of Panpsychism* Routledge.

Crane, Tim (2001). ‘Elements of Mind: An Introduction to the Philosophy of Mind.’ New York: Oxford University Press.

Cumming, Sam, "Names", The Stanford Encyclopedia of Philosophy (Spring 2013 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/spr2013/entries/names/>.

Davidson, D. (1970). ‘Mental Events’, in Davidson, D. ‘Essays on Actions and Events.’ Oxford: Oxford University Press, 207–223.

Davidson, D. (1987). 'Knowing One's Own Mind', in Proceedings and Addresses of the American Philosophical Association 60 (3): 441-458.

Descartes, Rene, (translated 1968 by F.E. Sutcliffe). ‘Discourse on Method and the Meditations’. London: Penguin.

Dennett, Daniel C. (1978). ‘Where am I?’ In Brainstorms. MIT Press.

Edgington, Dorothy (2004). ‘Two kinds of possibility’. Aristotelian Society Supplementary Volume 78 (1):1–22.

Farkas, Katalin (2008). ‘The Subject's Point of View’. Oxford University Press.

Feynman, R. (1994). ‘Six Easy Pieces’. New York: Basic Books.

Fine, Kit (1994). Essence and modality. Philosophical Perspectives 8:1-16.

Fodor, J. A. (1974). ‘Special sciences (or: The disunity of science as a working hypothesis)’. Synthese 28 (2):97-115.

Frege, G. (1952), “On Sense and Reference”, in P. Geach and M. Black, eds., Translations from the Philosophical Writings of Gottlob Frege, Oxford: Blackwell, pp. 56–79

Giberman, Daniel (2015). ‘Is Mereology a Guide to Conceivability?’ Mind 124 (493):121-146.

Godman, Marion, Mallozzi, Antonella & Papineau, David (forthcoming). ‘Essential Properties are Super-Explanatory: Taming Metaphysical Modality’

Goff, Philip (2012). ‘Does Mary know I experience plus rather than quus? A new hard problem’ Philosophical Studies 160 (2): 223-235

Goff, Philip & Papineau, David (2014). ‘What's wrong with strong necessities?’ Philosophical Studies 167 (3):749-762.

Goff, Philip, Seager, William and Allen-Hermanson, Sean, "Panpsychism", *The Stanford Encyclopedia of Philosophy* (Winter 2017 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/win2017/entries/panpsychism/>.

Hawthorne, John (2002). ‘Blocking definitions of materialism.’ Philosophical Studies 110 (2):103-13.

Heil, John. “Are we brains in a vat?”. In Heil, John (ed.) (2004), Philosophy of Mind: A Contemporary Introduction. Routledge.

Hempel, Carl G. (1980). ‘Comments on Goodman's ways of worldmaking.’ Synthese 45 (2):193 - 199.

Horgan, Terence E. (1984). Jackson on physical information and qualia. *Philosophical Quarterly* 34 (April):147-52.

T.H. Huxley and W.J. Youmans (1869). ‘The Elements of Physiology and Hygiene: A Text-Book for Educational Institutions.’ New York: Appleton & Co.

Jackson, Frank (1982). ‘Epiphenomenal qualia.’ Philosophical Quarterly 32 (April):127-136.

Jackson, Frank (1998a). ‘From Metaphysics to Ethics: A Defence of Conceptual Analysis.’ New York: Oxford University Press.

Jackson, Frank. (1998b), “Reference and Description Revisited,” in J. Tomberlin (ed.), Philosophical Perspectives 12: Language, Mind, and Ontology, Oxford: Blackwell, 201–218.

Jackson, Frank. (1998c), “Postscript on Qualia”, in Jackson: *Mind, Methods and Conditionals*, London: Routledge.

Jago, Mark (2016). Advanced Modalizing Problems. Mind 125 (499):627-642

Kant, Immanuel. *Critique of Pure Reason*, translated/edited by N Kemp Smith. Macmillan Press (1929).

Kaplan, D. (1978). “Dthat”. In Peter Cole (ed.), Syntax and Semantics. Academic Press, pp,221-243.

Kaplan, D. (1989), “Demonstratives/Afterthoughts”, in J. Almog, J. Perry and H. Wettstein, eds., Themes from Kaplan, Oxford: Oxford University Press, pp. 481–614.

Kim, Jaegwon (1989). The myth of non-reductive materialism. *Proceedings and Addresses of the American Philosophical Association* 63 (3):31-47.

Kim, Jaegwon (2003), “Mental Content”. In John Heil (ed.), *Philosophy of Mind: A Guide and Anthology*. Oup Oxford

Kment, Boris, "Varieties of Modality", The Stanford Encyclopedia of Philosophy (Winter 2012 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/win2012/entries/modality-varieties/>.

Kripke, Saul A. (1971). ‘Identity and necessity.’ In Milton K. Munitz (ed.), ‘Identity and Individuation.’ New York: New York University Press 135-164.

Kripke, Saul A. (1979), “A puzzle about belief”, in A. Margalit (ed.), Meaning and Use, Reidel. 239--83.

Kripke, Saul A. (1980). ‘Naming and Necessity.’ Cambridge, Massachusetts: Harvard University Press

Kripke, Saul A. (1982). ‘Wittgenstein on Rules and Private Language’. Harvard University Press.

LaPorte, Joseph, "Rigid Designators", The Stanford Encyclopedia of Philosophy (Summer 2011 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/sum2011/entries/rigid-designators/>.

Lau, Joe; Deutsch, Max, "Externalism About Mental Content", *The Stanford Encyclopedia of Philosophy (Winter 2012 Edition)*, Edward N. Zalta (ed.), URL =< http://plato.stanford.edu/archives/win2012/entries/content-externalism/>.

Lee, Geoffrey (2014). ‘Unity and essence in Chalmers' theory of consciousness.’ Philosophical Studies 167 (3):763-773.

Levin, Janet, "Functionalism", The Stanford Encyclopedia of Philosophy (Fall 2013 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/fall2013/entries/functionalism/>.

Levine, Joseph (2011). ‘The Character of Consciousness’, Review of David Chalmers, The Character of Consciousness, Oxford University Press, 2010. *Notre Dame Philosophical Reviews* 2011.

Lewis, David K. (1986). ‘On the Plurality of Worlds.’ Oxford: Blackwell.

Loar, Brian (1976), “The semantics of singular terms”. Philosophical Studies 30 (6):353 - 377.

Ludlow, Peter, "Descriptions", The Stanford Encyclopedia of Philosophy (Fall 2013 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/fall2013/entries/descriptions/>.

Lycan, William G. (2000), Philosophy of Language: A Contemporary Introduction. Routledge.

McGinn, Colin (1977). “Charity, interpretation, and belief,” *Journal of Philosophy*, 74: 521–535.

McGinn, Colin (1989). *Mental Content*. Blackwell.

McLaughlin, Brian and Bennett, Karen, "Supervenience", The Stanford Encyclopedia of Philosophy (Spring 2014 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/spr2014/entries/supervenience/>.

McTaggart, J. Ellis (1908). The unreality of time. *Mind* 17 (68):457-474.

Nagel, Thomas (1974). ‘What is it like to be a bat?’ Philosophical Review 83 (October):435-50.

Nichols, Shaun ; Pinillos, N. Ángel & Mallon, Ron (2016). Ambiguous Reference. Mind 125 (497):145-175.

Papineau, David (2001). ‘The rise of physicalism’, in Carl Gillett & Barry M. Loewer (eds.), ‘Physicalism and its Discontents.’ Cambridge: Cambridge University Press

Papineau, David (2002). ‘Thinking About Consciousness.’ Oxford: Oxford University Press.

Papineau, David (2016). 'Teleosemantics', in D.L. Smith (ed.), 'How Biology Shapes Philosophy: New Foundations for Naturalism.' Cambridge: Cambridge University Press.

Peacocke, Christopher (1993). “Externalist explanation,” *Proceedings of the Aristotelian Society*, 93: 203–230.

Perry, John (2009). ‘Subjectivity’, in Brian McLaughlin, Ansgar Beckermann & Sven Walter (eds.), ‘The Oxford Handbook of Philosophy of Mind.’ Oxford: Oxford University Press.

Priest, Graham, Tanaka, Koji and Weber, Zach, "Paraconsistent Logic", The Stanford Encyclopedia of Philosophy (Winter 2016 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/win2016/entries/logic-paraconsistent/>.

Putnam, Hilary (1973), “Meaning and reference”. *Journal of Philosophy* 70 (19):699-711.

Putnam, Hilary (1975), “The meaning of 'meaning'”. *Minnesota Studies in the Philosophy of Science* 7:131-193.

Putnam, Hilary, 1981, *Reason, Truth, and History*. Cambridge University Press.

Quine, W.V.O. (1951). Two Dogmas of Empiricism. Philosophical Review 60 (1):20–43.

Quine, W.V.O (1954) ‘Quantification and the Empty Domain’, Journal of Symbolic Logic, 19(3): 177–179

Quine, W.V.O. (1960). ‘Word and Object.’ Cambridge, Massachusetts: The MIT Press.

Reimer, Marga, "Reference", The Stanford Encyclopedia of Philosophy (Spring 2010 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/spr2010/entries/reference/>.

Robb, David and Heil, John, "Mental Causation", *The Stanford Encyclopedia of Philosophy* (Summer 2019 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/sum2019/entries/mental-causation/>.

Robinson, Howard, "Dualism", The Stanford Encyclopedia of Philosophy (Winter 2012 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/win2012/entries/dualism/>.

Russell, B. (1905), ‘On Denoting,’ Mind, 14: 479–93.

Schroeter, Laura, "Two-Dimensional Semantics", The Stanford Encyclopedia of Philosophy (Winter 2012 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/win2012/entries/two-dimensional-semantics/>.

Searle, J. (1958), “Proper Names”, Mind, 67(266): 166–73.

Searle, J. (1983), Intentionality, Cambridge: Cambridge University Press.

Segal, Gabriel (1999). *A Slim Book on Narrow Content*. MIT Press.

Shapiro, Stewart, "Classical Logic", The Stanford Encyclopedia of Philosophy (Winter 2013 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/win2013/entries/logic-classical/>.

Soames, Scott (2005). Reference and Description: The Case Against Two-Dimensionalism. Princeton: Princeton University Press.

Stalnaker, Robert (1978). ‘Assertion’, in P. Cole (ed), ‘Pragmatics’. New York: New York Academic Press, vol 9:315-332.

Stoljar, Daniel, "Physicalism", The Stanford Encyclopedia of Philosophy (Fall 2009 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/fall2009/entries/physicalism/>.

Strawson, P.F. (1950), “On Referring,” Mind, 59: 320–334

Stubenberg, Leopold, "Neutral Monism", The Stanford Encyclopedia of Philosophy (Fall 2014 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/fall2014/entries/neutral-monism/>.

Taylor, Kenneth A. (1989), “Narrow Content Functionalism and the Mind-Body Problem,” *Noûs*, 23: 355-372.

Tooley, Michael (1990). Causation: Reductionism versus realism. Philosophy and Phenomenological Research 50:215-236.

Uzquiano, Gabriel, "Quantifiers and Quantification", The Stanford Encyclopedia of Philosophy (Winter 2018 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/win2018/entries/quantification/>.

Walker, Ralph Charles Sutherland (1989). “The Coherence Theory of Truth: Realism, Anti-Realism, Idealism” London: Routledge.

Yablo, Stephen., 1992, “Mental Causation”, *Philosophical Review*, 101: 245–80.

Yablo, Stephen (2002). “Coulda, woulda, shoulda”. In Tamar S. Gendler & John Hawthorne (eds.), Conceivability and Possibility. Oxford University Press. pp. 441-492.