Chapter 4 – The Conceivability Argument

In the last chapter we discussed one of the two ‘big’ arguments against physicalism: the knowledge argument. In this chapter, we discuss the other: the conceivability argument. I will argue against the most well-known contemporary form of the conceivability argument: David Chalmers’ Two-Dimensional Conceivability Argument. I will then go on to present an alternative conceivability argument: the Transparency Conceivability Argument.

A crucial premise of the Transparency Conceivability Argument is *Phenomenal Transparency*: the thesis that phenomenal concepts are transparent. Thus, the moral of this chapter is similar to the moral of the last chapter: the Conceivability Argument, just like the threat to physicalism lying behind the Knowledge Argument, has force only if we can justify Phenomenal Transparency. In the next chapter I will present an argument for Phenomenal Transparency itself. However, once we justify Phenomenal Transparency, a more straightforward anti-physicalist argument presents itself – the Transparency Argument – which I will also present and defend in the next chapter. Hence, the Conceivability Argument becomes redundant in the critique of physicalism.

As in the last chapter, my superficial target is *pure physicalism* – physicalism in conjunction with the thesis that fundamental reality can be exhaustively described in the mathematico-economic vocabulary of physics – rather than physicalism as such. In chapter 6 I show how the case against pure physicalism built in the first half of the book is also a case against physicalism in general.

Conceivability Arguments

Two conceivability arguments

A long standing way of arguing that physicalism cannot account for consciousness is via some kind of conceivability argument. Conceivability arguments involve a *conceivability*
**principle**: a principle asserting some kind of link between conceivability and possibility. The most straightforward form of conceivability principle asserts that anything that is conceivable is possible. Formulated in terms of the conceivable/possible truth of a sentence, we get the following principle:

\[
\text{Simple Conceivability Principle (SCP)} - \text{If a sentence is conceivably true then it's possibly true.}
\]

Anti-physicalist conceivability arguments use a conceivability principle to move from something which is plausibly conceivable to a possibility which is plausibly inconsistent with physicalism. The structure of anti-physicalist conceivability arguments, then, is as follows:

1. **Conceivability** – The argument begins with a claim about conceivability.
2. **From conceivability to possibility** – Via the conceivability principle, a move is made from conceivability to possibility.
3. **From possibility to actuality** – It is argued that the possibility that has been demonstrated is inconsistent with the truth of physicalism, and hence that physicalism is false.\(^1\)

The inspiration for this style of argument comes from Descartes, who argued in something like the following way:

**The Cartesian Conceivability Argument**

Premise 1 – I can clearly and distinctly conceive of my mind existing without a brain/body.

Premise 2 – Anything I can clearly and distinctly conceive of is possible.

\(^1\) I will not be discussing ‘explanatory gap’ arguments (Levine 1983, 2004), as I’m inclined to agree with David Papineau (1998) that identities don’t need explaining. Focus on the explanatory gap is a powerful intuitive starting point for the anti-physicalist intuition, but it seems to me to quickly collapse into one of the other arguments.
Conclusion 1 – Therefore, it’s possible for my mind to exist without a brain/body.

Premise 3 – If it’s possible for my mind to exist without a brain/body, then my mind cannot be identical with a brain/body.

Conclusion 2 – Therefore, my mind is not identical with a brain or a body.²

One problem with this argument is that, even if it is sound, it does not obviously refute physicalism. For the physicalist need not hold that the mind is identical with the brain, or that mental states are identical with physical states, so long as facts about the mind and its states are grounded in physical facts. The physicalist is obliged to hold that for each mental fact $M$ (e.g. the fact that Sarah is feeling anxious), there is some physical fact $P$ (e.g. the fact that Sarah is in brain state X), such that $M$ is constitutively grounded in $P$. But it is a further step to identify $M$ with $P$ (e.g. to identify the fact that Sarah is anxious with the fact that she is in brain state X).

Many physicalists hold that mental states are multiply realisable, that is to say they hold that, for any mental state $M$, there is a wide variety of physical states whose instantiation is sufficient to ground $M$. It could be for example that the physical state which grounds pain in a human being is very different from the physical state which grounds pain in an octopus. Indeed, although physicalists are obliged to hold that all actual mentality is grounded in physical facts, many allow that mentality could be grounded in non-physical facts: that there are (non-actual) possible worlds in which there are ghosts and angels whose mental lives are grounded in non-physical goings on. The physicalist who embraces multiple realisation can happily accept the conclusion of the above argument.³

² Descartes 1645/1996, sixth meditation.

³ The premise of the Cartesian argument above is not that minds can exist without bodies, but that my mind is not identical with a body or brain. A physicalist who is open to the possibility of my consciousness being uploaded onto a computer can accept this conclusion. I don’t deny that there may be ways of making the Cartesian argument work (in Goff 2010a, 2012a, 2014 I attempted to do this), nor that there are other ways of interpreting Descartes. But it would be distracting to follow these threads in an already complicated chapter.
Hence, in its contemporary guise, most associated with David Chalmers, the Conceivability Argument against physicalism involves zombies (roughly bodies without minds) rather than ghosts (roughly minds without bodies). The word ‘zombie’ is a technical term in contemporary philosophy of mind, and so it is important to distinguish ‘philosophical zombies’ from the ‘zombies’ we encounter in Hollywood movies. A philosophical zombie is, by definition, a physical duplicate of a human which lacks consciousness. You stick a knife in a zombie and it screams and runs away, but it doesn’t actually feel pain. When a zombie crosses the road it looks both ways waiting for the traffic to stop and then carefully crosses the road, and yet it has no visual or auditory experience of the world around it. Zombies behave just like us, because their bodies and brains are physically just like ours, and yet there is nothing that it is like to be a zombie.

In cases of multiple realisation, grounded facts need not necessitate their grounds. However, it is generally accepted that a grounded fact or entity is necessitated by the obtaining of its ground. If physical fact $P$ constitutively grounds the fact that there is pain, then a world in which $P$ is instantiated is already thereby a world in which there is pain. It follows that $P$ could not possibly be instantiated without it being the case that there is pain. We can sum this up with the following broadly accepted principle:

**Necessitation**: If fact/entity $X$ constitutively grounds $Y$, then necessarily if $X$ exists/obtains then $Y$ exists/obtains.

Thus, if all actual facts involving consciousness are grounded in the physical facts, then the physical facts necessitate the reality of consciousness. It follows that there couldn’t possibly be a world which is indiscernible from the actual world physically but in which there is no consciousness. But of course a world of zombies just is a possible world which is physically

---

4 Chalmers 1996, 2009. The term ‘zombie’ with this meaning was coined by Robert Kirk 1974a, 1974b.

5 The principle is generally put in terms of grounding rather than constitutive grounding. Endorsement of Necessitation is the norm, see for example Fine 2012, Rosen 2010, and Trogdon 2013b. There are, however, some who deny it, such as Montero 2013, Leuenberger 2014 and Skiles 2015.
identical to the actual world in every respect, but in which nothing is conscious. The mere possibility of a zombie world, therefore, is inconsistent with physicalism.

Let us represent the complete pure physical truth about the actual world, i.e. the complete physical truth in so far as it can be specified in the mathematico-nomic vocabulary of physics, with the letter ‘P’. P specifies the pure physical properties and relations (i.e. those physical properties and relations that can be captured in the mathematico-nomic vocabulary of physics) of every fundamental physical entity throughout the whole of space and time, as well as the fundamental laws governing those entities. We can then form a sentence asserting that the world is a zombie world by conjoining P with the sentence ‘nothing is conscious’. Starting with the Simple Conceivability Principle, we can put the bare bones of the zombie conceivability argument as follows:

**The Simple Zombie Conceivability Argument**

*Premise 1* – ‘P & nothing is conscious’ is conceivably true.

*Premise 2* – *Simple Conceivability Principle (SCP)* – If a sentence is conceivably true then it’s possibly true.

*Conclusion* – Therefore, ‘P & nothing is conscious’ is possibly true.

*Premise 3* – If ‘P & nothing is conscious’ is possibly true, then consciousness is not grounded in the pure physical facts and hence pure physicalism is false.

*Conclusion 2* – Therefore, pure physicalism is false.

In fact we do not need the possibility of a whole world of full blown zombies in order to refute physicalism. Pure physicalism is an ambitious thesis: it is the view that all properties and states of affairs are grounded for in the pure physical facts. If there is even one actually instantiated conscious state which is not grounded in the pure physical facts, pure physicalism is false.

Therefore, taking the letter ‘Q’ to represent an arbitrary conscious state that is actually instantiated, we can give a more general conceivability argument as follows:

**The Simple Conceivability Argument**

*Premise 1* – ‘P and nothing has Q’ is conceivably true.
**Premise 2 – Simple Conceivability Principle (SCP)** – If a sentence is conceivably true then it’s possibly true.

**Conclusion 1** – Therefore, ‘P and nothing has Q’ is possibly true.

**Premise 3** – If ‘P and nothing has Q’ is possibly true, then Q is not grounded in the pure physical facts and hence pure physicalism is false.

**Conclusion 2** – Therefore, pure physicalism is false.

**Clarifying conceivability**

In general, Chalmers thinks of conceivability as *negative ideal conceivability*, and this is also the notion of conceivability I will work with throughout the book. There are two aspects of this notion of conceivability: the ‘negative’ bit and the ‘ideal’ bit.

A sentence S is negatively conceivably true just in case it’s not a priori that S is false: we cannot know a priori that ‘Water is made of ectoplasm’ is false, and in this sense its truth is negatively conceivable. In contrast, we can know a priori that the sentence ‘My Uncle John is a married bachelor’ is false, and hence the truth of this sentence is not negatively conceivable.

Chalmers also considers a positive notion of conceivability, which would involve positively imagining a situation in which the relevant sentence is true. In some ways positive conceivability fits better with our intuitive notion of conceivability. When reading Descartes’ *Meditations* one finds oneself being persuaded that in some sense one can positively imagine one’s mind existing in the absence of one’s body. However, positive conceivability is a tricky notion to define, as it requires us to get clear on the nature of imagination. In this context it would be unwise to focus on a notion of imagination closely tied to human mental faculties, as we have no reason to think that the limits of human faculties have any bearing on the limits of possibility. Conceivability understood as some kind of rational coherence is

---

6 I am here using ‘imagination’ in an intuitive sense, in contrast to Descartes’ technical sense in which it is tied to sense experience.
more likely to serve as a guide to possibility, and negative conceivability seems to capture this idea.

Similarly, it seems unlikely that the contingent limits of human reason have any bearing on genuine possibility; ideal conceivability abstracts away from such limits. Ideal conceivability contrasts with prima facie conceivability, where sentence S is prima facie (negatively) conceivable true just in case one can’t a priori rule out the truth of S upon initial reflection. Square circles are not prima facie conceivable, as it is immediately obvious that they can be ruled out a priori. But consider the following sentence ‘There are cousins, but no siblings.’ Upon first reflection, one might think that this proposition describes a coherent state of affairs. However, when one thinks about it a bit longer, it is clear that this is a state of affairs which reasoning can rule out, as one’s cousin is just the child of one’s parent’s sibling.

Complex mathematical falsehoods, e.g. ’57 x 89 = 5074’, are more straightforward examples of propositions that can be ruled out a priori but only after a certain amount of reflection. Time travel scenarios in which the past is changed, such as we find in the film Back to the Future, may also be incoherent in a subtle way.

A proposition is ideally (negatively) conceivable when no amount of a priori reasoning could rule out its truth. The sentence ‘Water is made of ectoplasm’ is thus ideally (negatively) conceivable: I can sit in my armchair thinking about it for as long as I like, but I’m not going to be able to work out that water is not made of ectoplasm. The sentence ‘There are cousins, but no siblings’, and the sentence ‘57 x 89 = 5074’, however, are not ideally negatively conceivable, as sufficient reasoning reveals them to be false.

---

7 If we’re worried about how to define idealised rational powers, we can define ideal negative conceivability as follows: Proposition P is ideally negatively conceivable just in case there is no possible world in which someone rules out the truth of P a priori, in virtue of understanding P, and without any empirical information other than what is required in order to understand P.
Type-A and type-B Physicalism

The impact of the Kripke/Putnam revolution on the mind-body problem

We can plausibly think of the zombie conceivability argument as the contemporary analogue of Descartes’ conceivability argument. But between Descartes and David Chalmers there was a fundamental revolution in analytic philosophy, associated with the work of Saul Kripke and Hilary Putnam, which made the defence of any kind of conceivability argument much less straightforward.¹

Kripke and Putnam persuaded most philosophers in the analytic tradition of the existence of a posteriori necessities: sentences which are necessarily true despite not being knowable a priori. One much discussed example is the sentence ‘Water is H₂O’. We cannot find out sitting in the armchair that water is essentially made up of molecules composed of two hydrogen atoms and one oxygen atom: we have to go out and observe the world to discover this. And yet, according to Putnam and Kripke, there is no possible world in which water fails to be H₂O.

Consider the XYZ world, a possible world just like ours in every respect except that the colourless, odourless stuff which fills oceans and lakes and falls from the sky has some chemical composition other that H₂O, call it ‘XYZ’. Is this a possible world in which water is XYZ rather than H₂O? Not according to Kripke and Putnam. Water is the actual colourless, odourless stuff that fills oceans and lakes and falls from the sky; it is this stuff (I am currently pointing at the water in my glass). But the actual colourless, odourless stuff in oceans and lakes (this stuff) is H₂O. The stuff in the XYZ world is merely stuff that superficially resembles water.

If there are a posteriori necessities, then SCP is false. An a posteriori necessity is conceivably false – we cannot know a priori that water is H₂O rather than XYZ – and yet not possibly false – water is H₂O in all possible worlds. Defenders of conceivability arguments must

either reject the existence of a posteriori necessities, or come up with some more nuanced version of the conceivability principle. In what follows I will go along with the general philosophical consensus that there are a posteriori necessities, and hence explore the possibility of more nuanced conceivability principles.

Before the Kripke-Putnam revolution, physicalists generally took themselves to be obliged to adopt some form of analytic functionalism: the view according to which all mental concepts are functional concepts: concepts which pick out mental states in terms of the causal role they play with respect to sensory inputs, behavioural outputs and other mental states. For the analytic functionalist, to suppose that someone is, say, in pain, is just to suppose that she has an inner state that plays the causal role we associate with pain, i.e. causing avoidance behaviour as the result of bodily damage, disposing the individual to try to make it stop, etc. If we can locate in physical reality states which play the relevant causal roles, then we can account for mentality in physical terms.⁹

One can understand why people might have taken this to be the only way to account for mentality once pure physicalism is assumed. If pure physicalism is true, then at a fundamental level our world is pure causal structure, i.e. can be completely described in a mathematico-nomic vocabulary. But causal structure can ground only more causal structure, and hence a world in which fundamentally there is only causal structure is a world in which there is nothing but causal structure. Therefore, if pure physicalism is true, the only way to analyse mentality in physical terms is to analyse it in causal terms.

The problem with this old-school form of physicalism is that it seems to deny the existence *phenomenal concepts*, the distinctive kind of concept we employ when we think about a

---

⁹ Ryle 1927, Putnam 1960, 1967, Smart 1959, Lewis 1966, 1970, Armstrong 1968. The situation is slightly more complex than I’m describing. Smart, Lewis and Armstrong defended versions of the ‘identity theory’, according to which mental states are identical with physical states. However, they held that the concepts we use to pick out those states are functional. I will overlook these subtleties, as they don’t directly bear on the arguments we will be considering.
conscious state in terms of what it’s like to have it. We can of course think of conscious states \textit{in a third-person way}: I can think of my anxiety in terms of what it’s caused by, and what it causes me to do. But I can also think of my anxiety \textit{in a first-person way}: in terms of what it’s like for me when I have it, in terms of how it feels; and intuitively this is very different kind of concept. The analytic functionalist is obliged to either deny that there are such first-person ways of thinking about our conscious states, or to analyse, somewhat implausibly, first-person ways of thinking about anxiety in terms of third-person ways of thinking about anxiety.

Post the Kripke-Putnam revolution, physicalists slowly started to realise that they had another option. Instead of denying the existence of phenomenal concepts, the physicalist might instead hold that they feature in a posteriori identities with physical or functional concepts. Just as water is a posteriori identical with H$_2$O, so anxiety is a posteriori identical with brain state X. Just as it is conceivable but not possible that the glass in front of me contains water whilst at the same time lacking H$_2$O, so it is conceivable but not possible that the creature in front of me instantiates brain state X without experiencing anxiety. Even if we cannot through armchair reasoning give \textit{a priori} analyses of facts about consciousness in terms of facts about causal structure, it may nonetheless be possible to empirically identify facts about consciousness with facts about causal structure.\textsuperscript{10}

Roughly based around this pre-Kripke/Putnam and post-Kripke/Putnam division of physicalists, Chalmers distinguishes between \textit{type-A physicalists} and \textit{type-B physicalists}.\textsuperscript{11} Type-A physicalists take there to be an a priori entailment from the physical truths to all other truths, including the \textit{phenomenal truths}, i.e. the truths about consciousness. That is to


\textsuperscript{11} Chalmers 2002.
say that if you knew enough about the workings of my body and brain you could in principle work everything there is to know about my conscious experience; you could work out for example whether or not I’m currently feeling anxious. For the type-A physicalist, zombies are inconceivable.

Type-B physicalists in contrast deny that the phenomenal truths can be deduced from complete knowledge of the physical truths; no matter how much you knew about my brain states and what they do, you would never be able to deduce whether or not I’m feeling anxious. Despite this epistemic difference between type-A and type-B physicalists, as forms of physicalism both are united in holding that the fact that I am anxious, and all other phenomenal facts, are constitutively grounded in the physical facts. And if the phenomenal facts are constitutively grounded in the physical facts, then zombies, whether or not they are conceivable, are impossible.

Since philosophers realised that type-B physicalism was an option, type-A physicalism has not been so popular. Whether or not there is a metaphysical gap between the pure physical facts and the facts about consciousness, it’s extremely plausible that there is an *epistemic gap* of the following form:

*The Epistemic Gap* – For any pure physical truth (i.e. a physical truth specified in a purely mathematico-nomic vocabulary) $C$, and any phenomenal truth (i.e. a truth concerning the instantiation of a conscious state, conceived of in terms of what it’s like to have it) $Q$, it’s not the case that $C$ a priori entails $Q$ (i.e. it’s not the case that one can rule out $<C\&\neg Q>$ a priori).

In the last chapter I argued that, although the knowledge argument does not refute physicalism, it does provide a fairly conclusive case for this epistemic gap. It follows from the epistemic gap that zombies are conceivable: if the phenomenal truths cannot be

\[\text{The Epistemic Gap}\]

\[\text{For any pure physical truth (i.e. a physical truth specified in a purely mathematico-nomic vocabulary) } C, \text{ and any phenomenal truth (i.e. a truth concerning the instantiation of a conscious state, conceived of in terms of what it’s like to have it) } Q, \text{ it’s not the case that } C \text{ a priori entails } Q \text{ (i.e. it’s not the case that one can rule out } \langle C\&\neg Q \rangle \text{ a priori).}\]

12 Ironically Frank Jackson (2007) is one of the few remaining proponents of type-A physicalism, having turned his back on poor Mary.
deduced from P – the complete pure physical truth about reality – then it will be conceivable that P is true but there is no consciousness. But in order to demonstrate the falsity of physicalism, we must move from the mere conceivability of zombies to their genuine possibility. In the next section I will explain David Chalmers’ strategy for trying to do this.

Moving from Conceivability to Possibility

The two-dimensional conceivability principle

As we have seen, the existence of a posteriori necessities, such as <water is H₂O>, is inconsistent with the Simple Conceivability Principle. In his full conceivability argument, therefore, Chalmers defends a more nuanced conceivability principle, which I shall call the ‘Two-Dimensional Conceivability Principle’, on account of its being defined in terms of his two-dimensional semantic framework, which I shall now introduce.¹³

We can think of the two-dimensional framework as a way of modelling what’s going on in Kripke’s Naming and Necessity, the book in which he argues that there are a posteriori necessities. Essentially Kripke’s claim is that, with respect to most scientific kinds, we can distinguish between the essence of the kind and its appearance property. For example, the essence of water is being made up of H₂O molecules; the essence of gold is being made up of atoms of atomic number 79.¹⁴ The appearance property of a kind, in contrast, is the property we use to pick it out. The appearance property of water is (roughly) being the colourless, odourless stuff that fills oceans and lakes, etc.; the appearance property of gold is

¹³ Chalmers 2009.

¹⁴ I am working with Kripke’s conception of the essence of water and gold. As we shall see in our discussion of Russellian monism from chapter 6 onwards, the concept <H₂O> is arguably opaque. However, I will continue to follow Kripke in talking of the essence of water as being H₂O for the sake of simplicity. We can think of ‘being H₂O’ as a placeholder for the essence of water whatever it turns out to be.
being heavy, yellowish, metal, etc. The two-dimensional framework captures this by distinguishing two aspects of meaning, one of which (the primary intension) corresponds to appearance, and one of which (the secondary intension) corresponds to essence.

Chalmers thinks of intensions as functions, from possible worlds to referents in the case of words, and from possible worlds to truth-values in the case of sentences. However, we can roughly capture primary and secondary intensions with descriptions and subordinate sentences, and for the sake of exegetical ease this is how I will think of them in what follows.

At the level of words, the primary intension is roughly captured by a description expressing the appearance property of the referent, and the secondary intension is captured by a description of referent’s essence. Thus the primary intension of ‘water’ is approximated by the description ‘The colourless, odourless stuff that fills oceans and lakes and falls from the sky, etc.’ (from now shortened to ‘the watery stuff’) and the secondary intension of ‘water’ is captured by ‘H₂O.’

For any sentence S, the primary/secondary intension is roughly captured by the subordinate sentence which results when the terms in S are replaced by descriptions which roughly capture the primary/secondary intensions of those terms. Thus, the primary intension of ‘Water is XYZ’ is approximated by the sentence ‘The watery stuff is XYZ’, whilst the secondary intension of ‘Water is XYZ’ is captured by the sentence ‘H₂O is XYZ’.

15 Crucially on Chalmers’ framework the primary intention function can be evaluated a priori, which is why it is appropriate to think of it as capturing an ‘appearance property’. If the property in virtue of which reference is fixed were not a priori accessible to the concept user, then the concept in question would lack a primary intension; such a concept would be ‘radically opaque’, to use the terminology developed below.

16 The worlds at which the primary intension are centred, which captures the role of indexicals in fixing reference. I will overlook this for the sake of simplicity.
The secondary intension of a word reflects the essence of its referent; the secondary intension of a sentence specifies the worlds at which the sentence is true or false. So for example, knowing that the secondary intension of ‘Water is XYZ’ is captured by the sentence ‘H\textsubscript{2}O is XYZ’ makes it apparent that ‘Water is XYZ’ is false at all possible worlds. The primary intension reflects how the meaning of a term or sentence is dependent on facts about the actual world, e.g. the term ‘water’ refers to whatever is the watery stuff in the actual world, which is reflected in its primary intension.

Let us begin to move towards the link Chalmers wants to make between conceivability and possibility using this framework. Consider again the XYZ world, the world at which the watery stuff is XYZ. The secondary intension of ‘Water is XYZ’ – i.e. ‘H\textsubscript{2}O is XYZ’ – is false at the XYZ world. However, the primary intension of ‘Water is XYZ’ – i.e. ‘the watery stuff is XYZ’ – is true at the XYZ world. Even though the secondary intension of ‘Water is XYZ’ and hence the overall sentence is false at every possible world, there is a genuine possible world at which its primary intension is true.

Chalmers takes this to be the case for every sentence which is (ideally negatively) conceivably true: even though it may be false at every possible world, there will be some genuine possible world at which its primary intension is true. Thus, whilst rejecting SCP on Kripkean grounds, Chalmers takes the following more nuanced principle linking conceivability and possibility to be true:

\textbf{2D-CP} – If a sentence is conceivably true, then its primary intension is true at some possible world.

With this principle in place, we can now apply it to the general conceivability argument. Assuming as I am that the pure physical truths do not a priori entail the phenomenal truths, the sentence ‘P and nothing has Q’ is conceivably true. And hence if we accept 2D-CP we can infer that its primary intension is true at some possible world. We have inferred from the conceivability of the pure physical facts without Q to a genuine possibility. But we have yet to show that this possibility is inconsistent with pure physicalism. We have already established that pure physicalism is inconsistent with the possible truth of the whole
sentence ‘P and nothing has Q’, but the whole sentence is possibly true only if its secondary intension is possibly true; so far we have only the possible truth of its primary intension.

At this point, Chalmers draws on another aspect of Kripke’s work. As discussed above, Kripke’s insight in *Naming and Necessity* is that, with many scientific kinds, the property we use to think about the kind (the appearance property) comes apart from its essence. Water is essentially H₂O, but we think about it in terms of one of its non-essential properties: being the watery stuff. In this sense our thought about water is indirect. However, Kripke goes on to deny that this is the case when it comes to consciousness. We think about a pain in terms of how it feels, but how pain feels is not a non-essential way pain happens to appear to us: pain just is a feeling. In contrast to thought about water, thought about pain is direct. In general, we think about a conscious state in terms of what it’s like to be in it; in each case there is no appearance property to be distinguished from the essence of the conscious state. Call this Kripkean thesis that phenomenal concepts do not pick out conscious states in terms of non-essential appearance properties of those states the ‘Direct Reference Thesis’.

The primary intension of a word specifies the appearance property of its referent, the secondary intension its essence. Hence, on the two-dimensional framework, the Direct Reference Thesis comes out as the thesis that terms expressing phenomenal concepts lack distinct primary and secondary intensions. If terms expressing phenomenal concepts lack distinct primary and secondary intensions, then sentences involving those terms lack distinct primary and secondary intensions (unless of course they also involve other terms which have distinct primary and secondary intensions). And if the sentence ‘P & nothing has Q’ lacks distinct primary and secondary intensions, and we know its primary intension is possibly true, then we can infer that its secondary intension is possibly true.

We can state Chalmers’ version of the conceivable argument as follows:

**Two-Dimensional Conceivability Argument**

*Premise 1* – ‘P and nothing has Q’ is conceivably true.

*Premise 2* – 2D-CP: If a sentence is conceivably true, then its primary intension is true at some possible world.
**Conclusion 1** – The primary intension of ‘P and nothing has Q’ is true at some possible world.

**Premise 3** – It is not the case that ‘P and nothing has Q’ has a primary intension which differs from it secondary intension (Direct Reference Thesis).\(^{17}\)

**Conclusion 2** – The secondary intension of ‘P and nothing has Q’ is true at some possible world, and hence physicalism is false.

As Chalmers admits, there is gap in his justification for premise 3. The Direct Reference Thesis has been offered in defence of the claim that terms expressing phenomenal concepts lack distinct primary and secondary intensions. However, the sentence ‘P and nothing has Q’ also contains physical concepts: those used to specify P. We have not yet given reason to think that physical concepts lack distinct primary and secondary intensions, and hence have incomplete justification for the claim that ‘P and nothing has Q’ lacks distinct primary and secondary intensions.

However, if we take this to be an argument against pure physicalism then the problem goes away, as pure physical facts can (by definition) be exhaustively captured in a purely mathematico-nomic vocabulary, and causal and formal concepts plausibly do not have distinct primary and secondary intensions. When we think about a causal property in terms

\(^{17}\) Chalmers does argue (Chalmers 2009: 153) that that we can run the argument without assuming that phenomenal concepts lack distinct primary and secondary intensions, by focusing on the primary intensions of phenomenal concepts. However, this move is also reliant on the assumption that every concept has a primary intension which affords a substantive a priori grasp of some significant property its referent, either its essence or a property which uniquely identifies it in the actual world. Assuming this, even if phenomenal concepts do not afford us a transparent grasp of the essence of consciousness properties themselves, they will afford us a transparent grasp of the properties we use to pick out conscious properties. Hence, this alternative way of running the argument is also subject to the objection I will make to Chalmers’ main argument: that he adopts without sufficient argument a framework which rules out from the get go the dominant form of type-B physicalism according to which phenomenal concepts are radically opaque.
of its causal nature, or a mathematical property in terms of its mathematical nature, we pick out the property directly rather than in terms of an accidental appearance property.

This is why matters are simplified if we focus in the first instance on pure physicalism rather than physicalism in general. In the next chapter I will move on to the refutation of physicalism in general, but for the moment I will interpret Chalmers’ argument as an attempted refutation of pure physicalism.

**Against the two-dimensional conceivability principle**

The two-dimensional framework is built on the assumption that each term has a primary intension, which expresses the appearance property of its referent: the property in virtue of which reference is determined, and which is a priori accessible to concept user. To put it another way: the concept expressed by each term reveals, i.e. renders a priori accessible, a feature of the referent which uniquely identifies it in the actual world. Where the primary and secondary intensions are identical the property revealed uniquely identifies the referent in all possible worlds; but even when primary and secondary intentions are distinct, a substantive feature of the referent is a priori accessible.\(^{18}\)

It is crucial to appreciate that the adoption of this framework rules out from the get go the dominant form of type-B physicalism. Most type-B physicalists agree with the Kripkean Direct Reference Thesis: that we do not pick out conscious states in terms of non-essential appearance properties. But type-B physicalists also tend to deny that phenomenal concepts reveal essential features of conscious states. The standard type-B line is that phenomenal concepts reveal *no* substantive features of their referents.

We can make this point clearer by returning to the distinction between transparent and opaque concepts discussed in chapters 1 and 3:

\(^{18}\) I am over-simplifying slightly: in fact Chalmers thinks indexicals are also involved in determining reference, hence the need for the worlds at which primary intensions are evaluated to be centred (cf. footnote 14).
• *Transparent Concept* – A concept C referring to entity E is transparent just in case C reveals the nature of E (i.e. what it is for E to be part of reality is a priori accessible for someone possessing C, in virtue of possessing C), e.g. <sphericity>, <party>.

• *Opaque Concepts* – A concept C referring to entity E is opaque just in case C reveals little or nothing of the nature of E, e.g. <water>, <gold>.

We can further distinguish between two kinds of opaque concept:

• *Mildly Opaque* – A mildly opaque concept is one that reveals (under a transparent concept) *significant accidental properties* of its referent, typically those which uniquely identify it in the actual world.\(^{19}\)

• *Radically Opaque* – A radically opaque concept reveals *no significant properties* (neither essential nor accidental) of its referent.\(^{20}\)

We can think of radically opaque concepts as ‘blind pointers’, allowing the concept user to latch onto an entity without conceiving of it in terms of any of its features.

As previously noted, logical or mathematical concepts are plausibly examples of transparent concepts. For example, the essence of sphericity is a priori accessible: if you possess the concept of sphericity, and you’re clever enough, you can work out that for there to be

\(^{19}\) To be more precise, it would be better to say that the significant property typically revealed uniquely identifies the referent in a *centred* possible world considered as actual, where the centre picks out the spatio-temporal location of the speaker. This device models the contribution of indexicals in fixing reference. For more on this see Chalmers 2004.

\(^{20}\) Putting the word ‘significant’ into the definition of a radically opaque concept adds an element of imprecision. All concepts reveal trivial properties of their referents, e.g. self-identity. And probably all concepts reveal the general ontological category of their referents: in the case of <water> we arguably know a priori that the referent is *uncountable stuff* rather than a countable entity. Perhaps cleverer people than myself might be able to define the notion more precisely, but I don’t think it is essential to do this in order for this category of concept to be useful.
something spherical is for there to be something with all its surface points equidistant from its centre. Most ordinary language natural kind concepts are plausibly examples of mildly opaque concepts: we conceive of water, for example, as the watery stuff. What about radically opaque concepts? How can we possibly think about something, without thinking about it in terms of any of its properties?

Type-B physicalist David Papineau draws on another aspect of Kripke’s work to try to make sense of this: his causal theory of names (it is ironic that both sides of the contemporary mind-body debate draw on Kripke’s work). According to the descriptivist theory of names popular before Kripke, names refer in virtue of associated descriptions. It might be supposed, for example, that we pick out Shakespeare as ‘the person who wrote Romeo and Juliet, Hamlet, Macbeth, etc.’ On the descriptivist theory, names are mildly opaque: we conceive of Shakespeare in terms of an accidental property of being the writer of such and such plays, which uniquely identifies him in the actual world.

Kripke’s objection to such an account is that, for any such description associated with a name, it is coherent to suppose that the individual named does not in fact satisfy that description. It could turn out, as indeed some have claimed, that Shakespeare did not write the plays he is associated with. This makes no sense if ‘Shakespeare’ just means ‘the person who wrote Romeo and Juliet, Hamlet, Macbeth, etc.’ If the descriptivist theory of names is true, then the view that Shakespeare did not write Hamlet would not be historically suspect but incoherent.

After rejecting the descriptivist theory of names on these grounds, Kripke argues for the causal theory of names as its replacement. On this view, there is a complicated causal chain stretching back through history which connects the term ‘Shakespeare’ to Shakespeare himself, and it is in virtue of this causal chain that ‘Shakespeare’ refers to Shakespeare. Plausibly this chain began with Shakespeare’s parents baptising him ‘William Shakespeare’, and the name was then passed on from person to person. The existence of such a chain

---

does not depend on any specific descriptive information being associated with Shakespeare, and hence the term ‘Shakespeare’ does not end up being a priori associated with any substantive property of its referent. Reference is determined ‘outside of the head’ rather than via the understanding of the concept user. On this view names are radically opaque: it is causal facts that hook up the concept user to Shakespeare, rather than anything the concept user grasps about Shakespeare’s accidental or essential nature.

The standard view of contemporary physicalists is that phenomenal concepts are also radically opaque. A phenomenal concept is a certain kind of evolved sub-personal capacity for referring to a certain inner state of the concept user. The reference of this concept is determined either by its evolved function, or by certain causal connections between the concept and the inner state it tracks. Reference fixing does not involve the understanding of the concept user, and hence deployment of the concept does not essentially involve any kind of substantive grasp of the referent; we merely have a capacity to ‘blindly point’ at certain inner states. The inner states blindly pointed at turn out to be neurophysiological or functional states of the brain.

Some type-B physicalists are quite explicit about this. Brain McLaughlin says the following:

Phenomenal concepts are of two sorts: nondescriptive name concepts and type-demonstrative concepts; as such, phenomenal concepts lack any descriptive content. They do not conceptually reveal anything about the essential nature of phenomenal properties: they simply name or demonstrate them.\(^{22}\)

Type-B physicalist David Papineau goes further in asserting that all basic concepts are radically opaque:

No doubt there are ways of thinking of things that make certain essential properties a priori knowable. But I take such a priori knowledge to derive from (possibly implicit) compositionality in the relevant modes of thinking, and so not to be

\(^{22}\) McLaughlin 2001: 324.
associated with the most basic ways in which thought makes contact with reality...When it comes to these basic points of contact, I find it hard to take seriously any alternative to the assumption that our atomic concepts are related to reality by facts external to our a priori grasp, such as causal or historical facts.... I don’t recognise any way in which the mind ‘captures’ something, apart from simply referring to it.23

Whilst other type-B physicalists may not be explicit about their commitment to Radical Phenomenal Opacity – the thesis that phenomenal concepts are radically opaque – it is often implicit in their favoured theories of phenomenal concepts as demonstratives,24 indexicals,25 recognitional concepts,26 or concepts which refer in virtue of facts about teleology or causal connections.27

Radical Phenomenal Opacity provides a nice explanation of the conceivable impossibility of zombies. A phenomenal concept of <pain> ‘blindly points’ at brain state X. Pain, the thing referred to by the phenomenal concept <pain>, just is brain state X, and hence pain and brain state X could not possibly exist without each other. But because this is blind pointing, it is not a priori that <pain> denotes brain state X, and so it is conceivable for the concept user that something could have brain state X without feeling pain. Zombies are conceivable but impossible.

Return to the Direct Reference Thesis. Kripke expresses it thus:

Pain...is not picked out by one of its accidental properties; rather it is picked out by the property of being pain itself, by its immediate phenomenological quality. Thus

25 Tye 1995, chap. 6, Lycan 1996, sect. 3.3.
pain, unlike heat, is not only rigidly designated by ‘pain’ but the reference of the designator is determined by an essential property of the referent.\textsuperscript{28}

What did Kripke mean here? As I read this, Kripke is saying that when we reflect on pain we conceive of it in terms of its essential nature, the view I will defend in chapter 5. But that’s not how type-B physicalists interpret Kripke’s thesis. In Brian Loar’s classic defence of the view that became known as ‘type-B physicalism’, he talks of our ‘grasping the essence’ of phenomenal qualities. However, on closer examination we find that all he really meant by this is that we don’t think about phenomenal qualities in terms of their accidental properties, i.e. that phenomenal concepts are not mildly opaque:

Phenomenal concepts, as we have seen, do not conceive of their reference via contingent modes of presentation. And so they can be counted as conceiving phenomenal qualities directly. Calling this a grasp of essence seems to me all right, for phenomenal concepts do not conceive their references by way of their accidental properties . . . \textsuperscript{29}

If, as is standard among type-B physicalists,\textsuperscript{30} we take the Direct Reference Thesis to be equivalent to the thesis that phenomenal concepts are not mildly opaque, then Radical Phenomenal Opacity entails that thesis; according to Radical Phenomenal Opacity phenomenal concepts are not mildly opaque but radically opaque. But Radical Phenomenal Opacity also entails the falsity of premise 2 of the 2D-Conceivability Argument, the crucial 2D-CP allowing us to move from conceivability to possibility. If phenomenal concepts lack a primary intention, then ‘P and nothing has Q’ lacks a primary intension, and hence ‘P and nothing has Q’ does not have a primary intension that’s true at some possible world, despite its conceivability. 2D-CP has a shot at being true only if each concept has a primary intension; that is to say, only if there are no radically opaque concepts.

\textsuperscript{28} Kripke 1972: 15-3.

\textsuperscript{29} Loar 1990/1997: 105.

None of this is to say that Radical Phenomenal Opacity is in itself a plausible thesis. It seems to me manifestly an implausible view about phenomenal concepts, and I hope it strikes the reader that way too. Surely when I attend to my pain introspectively, when I think about it in terms of how it feels, I understand something about its nature. That’s why I care when people feel pain. Judgements of the moral significance of pain are difficult to make sense if pain is just that thing I know not what. In knowing how pain feels, I know something of what pain is, and it’s this understanding which grounds my awareness that pain is all things being equal a bad thing. Type-B physicalists often boast that, in contrast to analytic functionalism, their view captures our intuitive view about phenomenal concepts. But the view that phenomenal concepts are blind pointers seems to me just as revisionary as the view that phenomenal concepts admit of a priori causal analysis.\(^{31}\)

Nonetheless, if we dislike Radical Phenomenal Opacity, we need to argue against it, or at least assert its rejection as an explicit premise in our arguments. It is dialectically inappropriate to just rule it out by the way we set up the discussion; but this is exactly what Chalmers does in framing the argument in terms of the two-dimensional framework. Moreover, Chalmers rules out Radical Phenomenal Opacity by means of a very wide-ranging, highly contentious semantic assumption. The two-dimensional framework assumes that every single concept has a primary intension, and hence that no concepts are radically opaque. It ought to be possible for an anti-physicalist not to stick out her neck so far: rejecting Radical Phenomenal Opacity, but remaining agnostic on whether other concepts, such as names, are radically opaque.

If Chalmers had no argument for adopting a radically-opaque-concept-banning framework, then setting things up in this way would wildly beg the question against the physicalist. As it is, he ultimately defends the two-dimensional framework on the grounds that it allows us to defend a connection between conceivability and possibility, in spite of the Kripkean

\(^{31}\) I made this argument in Goff 2011.
exceptions to SCP. However, in the next section I will show how a more straightforward link between conceivability and possibility can be maintained, and one which does not involve the contentious semantic assumption that there are no radically opaque concepts. In the light of this, Chalmers’ argument for the adoption of the two-dimensional framework, and consequently the Two-Dimensional Conceivability Argument itself, fails.

The Transparency Conceivability Principle

When conceiving of entity e under a transparent concept one understands (or has a priori access to) the nature of e: one understands what it is for that e to be part of reality (to exist if e is an object, to be instantiated if e is a property, to take place if e is an event, etc.). At the level of whole thoughts, the analogue of the transparent concept would be a thought such that in thinking it one understands (or has a priori access to) what it would be for the possible fact being conceived of to obtain.

It is extremely plausible that each property and kind has a unique grasppable essence, which distinguishes it from all other properties and kinds. One can understand, for example, what it is for partyhood or sphericity to be instantiated. It is much more contentious whether fundamental particulars have graspable essences. Of course particular individuals such as electrons fall under kinds, and we can perhaps understand what it is for each of those kinds to be instantiated. But this does not involve understanding what it is for electron X as

32 Chalmers argues that the link between conceivability and possibility allows us to think of metaphysical modality and epistemic modality as derived for a single set of possible worlds, and so avoid ‘modal dualism’ (see the ‘Modal rationalism’ section of Chalmers 2009). The real worry here I think is brute modal truths, which the position I end up with at the end of chapter 5 avoids. I explain in more detail in Goff & Papineau (2014) how my position, although perhaps strictly speaking a form of ‘modal dualism’, avoids the brute modal truths Chalmers worries about.

33 Platonists might take properties to be part of reality even if uninstantiated, in which case we could talk about it was is required for the property to have concrete reality.
opposed to electron Y to exist. I could perhaps know the essences of each of electron x’s properties, without knowing what it is for that particular electron, as opposed to some distinct qualitative duplicate, to exist. This does not seem to be because of cognitive limitations, because of something which, as it were, God understands but we don’t. Rather it is natural to suppose that fundamental individuals are brute one-offs: the particular existence of some particular fundamental individual is just not the kind of thing that can be understood. It is just a brute fact about electron X that it is the electron it is and not some other electron.34

If fundamental individuals are brute one-offs, then the notion of a transparent concept of an individual gets no purchase. A transparent concept reveals what it is for its referent to exist. If the particular existence of electron X is not the kind of thing that can be understood, then there could not be a transparent concept referring to X itself, as opposed to the properties and kinds which X instantiates. This does not entail that we cannot refer specifically to X rather than any other electron; we can perhaps do this by demonstration or in terms of a property which uniquely identifies X in the actual world. But such forms of reference are not transparent.

34 What about non-fundamental individuals? It is plausible that non-fundamental particulars are essentially defined in terms of their origins, and that their origins can be uniquely specified only with respects to the fundamental entities which ground the facts about their origins. If this is correct, then non-fundamental individuals are defined in terms of fundamental individuals, and to this extent their essences will inherit an element of bruteness from the bruteness of those fundamental individuals. A term referring to a non-fundamental individual e could be said to be ‘transparent’ in a secondary sense, if it reveals what it is for e to exist in terms of the fundamental entities that define its origin. No doubt the essence of many non-fundamental entities is radically vague, but this is no bar in principle to there being a concept which reveals that radically vague essence.
To the extent that we conceive of facts involving individuals that lack unique graspable essences, to that extent one cannot fully understand what it is for the fact being conceived of to obtain. For example, suppose that Jane lacks a unique graspable essence that distinguishes her from all other women. In this case one could understand what it is for something to be in pain, but one could not understand what it is for *Jane as opposed to Clare* to be in pain. This is not because there is something about the state of affairs of Jane’s being in pain which we can never get at; it is just that Jane, unlike her pain, is not the kind of thing whose concrete existence is comprehensible.

The situation seems at first paradoxical: we understand everything that can be understood about the fact that Jane is in pain, even though we don’t entirely understand the nature of that fact. But the paradox is removed when we again appreciate that there is something about the fact – the particular existence of Jane as opposed to any other woman – which is not the kind of thing that can be understood. Consider the following analogy: we can see everything that can be seen in the room without seeing the electrons in the room, as electrons are not the kind of thing that can be seen. Similarly, we can understand everything that can be understood about a particular fact without understanding everything about that fact, if there is an aspect of the fact that is not the kind of thing that can be understood.

For these reasons, thoughts involving singular concepts referring to individuals that lack graspable essences can never be entirely transparent. If we want to ensure that we are

---

35 Do things which don’t have graspable essences have ungraspable essences? Perhaps one might say that they do in the trivial sense that what it is for fundamental particular X to exist is just *for X to exist*, but I prefer to stipulate that essences are graspable, so that by definition we cannot form a transparent concept of something that lacks a graspable essence.

36 I don’t mean to say that it is obviously true that persons lack unique graspable essences. If a person can be defined in terms of its origins, then we can form a concept of a person that is transparent in the secondary sense given in footnote 34.
dealing with fully transparent thoughts, then we can focus on thoughts involving
quantification, such as <there are at least two things that feel pain>, rather than thoughts
involving singular concepts, such as <Jane feels pain>. We can give sufficient conditions for
a transparent thought and a transparent sentence (i.e. a sentence expressing a transparent
thought) as follows:

- A thought is transparent if (i) it involves no singular concepts, (ii) each of the
  property or kind denoting concepts involved in it are transparent.
- A sentence is transparent if (i) it involves no singular terms, (ii) the properties or
  kinds expressed by its terms are characterised by those terms under transparent
  concepts.

Having clarified what transparent thoughts are, I turn now to their relevance to the
relationship between conceivability and possibility. When thinking a transparent thought,
one thereby grasps what one is conceiving of, metaphysically speaking. In conceiving that
there are at least two spheres I plausibly understand what it is for that fact to obtain. In
conceiving that there is water, in contrast, I need not understand what it is for that fact to
obtain: finding this out requires empirical knowledge of the chemical composition of the
actual watery stuff.

Kripke’s examples of a posteriori necessities open up a gap between conceivability and
possibility. We can plausibly explain that gap in terms of the opacity of thought. We do not
have a priori access to what it is for there to be water that is XYZ, to the nature of the fact
that would have to obtain for that thought to be true. As a consequence, we are unable to
work out a priori that that fact is impossible, and hence the thought necessarily false. If we
knew that for there to be water that is XYZ is for there to be H₂O that is XYZ, it would be

37 I am assuming that the concepts <existence> and <identity> are transparent.

38 Even then I don’t think we’d grasp the essence of water, for reasons we will get to in chapter 6, but I’ll
overlook this subtlety here.
apparent that this fact obtains at no possible world.\textsuperscript{39} In other words, it is plausibly our ignorance of the nature of what we are conceiving of that opens up the Kripkean gap between conceivability and possibility.

If the Kripkean gap between conceivability and possibility is to be explained in terms of the opacity of the concepts involved in conceiving, then we need not take there to be such a gap when our thought is wholly transparent. In conceiving that there are two spheres, I understand what it would be for this fact to obtain. The Kripkean examples give me no reason to doubt that in this case I can infer from conceivability to possibility.

We may therefore, consistent with Kripkean a posteriori necessities, propose that conceivability entails possibility \textit{when you grasp what you're conceiving of (metaphysically speaking)}. We can sum up this view with the following alternative to Chalmers’ Two-Dimensional Conceivability Principle:

\textit{Transparency Conceivability Principle (TCP)} – If a transparent sentence is conceivably true, then it’s possibly true.

TCP is logically weaker than 2D-CP. 2D-CP says that every conceivably true sentence corresponds to some genuine possibility, whilst TCP talks only of a certain epistemologically very special group of sentences: the transparent ones. 2D-CP entails TCP – at least on the plausible assumption that a transparent sentence is a sentence with identical primary and secondary intensions – as both principles predict that every transparent sentence will be true at some possible world. TCP does not, however, imply 2D-CP: TCP has nothing to say about non-transparent sentences, and hence does not entail the things 2D-CP says about such sentences.

So 2D-CP goes beyond TCP in claiming that every conceivably true sentence corresponds to a genuine possibility. But it is able to say this only because of the implicit ban on radically opaque concepts. If there is a conceivably true sentence $S$ containing a radically opaque

\textsuperscript{39} I am assuming that the chemical definition of XYZ is inconsistent with the chemical definition of H$_2$O.
concept R, then R does not have a primary intension (as radically opaque concepts do not have primary intensions) and hence (in virtue of containing R) S does not have a complete primary intension, and hence S does not have a primary intension that is true at some possible world. Chalmers’ argument for his radically-opaque-concept-banning framework is that it allows us to maintain a link between conceivability and possibility in spite of Kripkean a posteriori necessities. However, TCP does the same thing without banning radically opaque concepts, and hence Chalmers has failed to provide us with a justification for a general ban on radically opaque terms and concepts.\textsuperscript{40}

We have not yet positively argued that there is \textit{any} link between conceivability and possibility. The point here is that the fact that we can link conceivability to possibility via TCP rather than the weaker 2D-CP undermines Chalmers’ defence of 2D-CP. Also undermined is the fairly common argument that the Kripke-Putnam examples of a posteriori necessity conclusively demonstrate that \textit{no} inference can be made from conceivability to possibility. For 2D-CP and TCP allow for inferences from conceivability to possibility, but are consistent with the Kripke-Putnam counterexamples to SCP.

Suppose we accept TCP. We could then use it to run the following form of the conceivability argument:

\textbf{Transparency Conceivability Argument}

\textit{Premise 1} – ‘P and nothing has Q’ is conceivably true.

\textit{Premise 2} – (TCP) If a transparent sentence is conceivably true, then it’s possibly true.

\textit{Premise 3} – ‘P and nothing has Q’ is transparent.

\textit{Conclusion} – ‘P and nothing has Q’ is possibly true, and so pure physicalism is false.

The Transparency Conceivability Argument is dialectically superior to the Two-Dimensional Conceivability Argument in that it doesn’t rely on a semantic framework which presupposes that the dominant form of type-B physicalism is false. However, whilst there is a broad

\textsuperscript{40} I make this criticism of Chalmers in Goff and Papineau 2014.
consensus on both sides of the debate that premise 3 of the Two-Dimensional Conceivability Argument is true, premise 3 of the Transparency Conceivability Argument is likely to be much more contentious. Let us examine this in more detail, setting on one side for the moment the question of whether TCP itself is true.

Whether or not ‘P and nothing has Q’ is transparent depends on whether or not the concepts involved in ‘P and nothing has Q’ are transparent, which obviously depends on whether pure physical concepts are transparent and on whether phenomenal concepts are transparent. It is extremely plausible that pure physical concepts are transparent. To know the causal role of a causal role property, or the mathematical description of a mathematical property, is to know its essence. The much more controversial question is: Are phenomenal concepts transparent?

Chalmers argues that phenomenal concepts are transparent on the basis of the Kripkean Direct Reference Thesis: the thesis that phenomenal concepts are not mildly opaque. If, as the two-dimensional framework assumes, all concepts are either transparent or mildly opaque, and phenomenal concepts are not mildly opaque, then it follows that phenomenal concepts are transparent. Hence, within the two-dimensional framework, the Direct Reference Thesis entails Phenomenal Transparency (the thesis that phenomenal concepts are transparent).

However, we have already seen that there is another option available to the physicalist: she can suppose that phenomenal concepts are radically rather than mildly opaque. The two-dimensional framework rules out this option from the outset, but without a good argument for adopting the two-dimensional framework this has little dialectical force.

We may also question the implicit assumption hitherto made that the categories of ‘transparent’ and ‘opaque’ are exhaustive. Perhaps some property/kind concepts reveal something but not everything of what is for their referents to have instances. It’s plausible to think the concept <being a sphere the same mass as the Earth> is an example of such a concept. One aspect of what it is for there to be a sphere the same mass of the Earth – there being something which is spherical – is transparently understood by the concept user; a full
understanding of what it is for there to be a sphere the same mass as the earth can be
known only through empirical investigation to find out the mass of the Earth.

Including such concepts, we can make our three-fold categorisation of concepts four-fold:

**Transparent** – A term/concept referring to K is transparent just in case it reveals the
essence of K.

**Translucent** – A term/concept referring to K is translucent just in case it reveals some
significant aspect of the essence of K, but not the complete essence.

**Mildly Opaque** – A term/concept referring to K is mildly opaque just in case it is not
transparent or translucent, but it transparently reveals (i.e. reveals under a
transparent concept) significant accidental properties of K, typically properties which
uniquely identify K in the actual world.41

**Radically opaque** – A concept referring to a property/kind K is radically opaque just in
case it does not transparently reveal any significant properties (neither essential nor
accidental) of K.

The type-B physicalist can straightforwardly avoid the transparency conceivability argument
by defending *Radical Phenomenal Opacity*, the thesis that phenomenal concepts are
radically opaque. Indeed if we have no a priori insight into the essential nature of
phenomenal properties, then it’s hard to see what philosophical grounds we could have for
denying that they are pure physical properties.

It is less clear whether the physicalist can avoid the TCP zombie argument by defending
*Phenomenal Translucency*, the thesis that phenomenal concepts are translucient. If a
property denoting concept C is translucient, then there is some aspect of its referent, call it

41 As noted in footnote 17, it would probably be better to say that the property revealed uniquely identifies
the referent in a *centred* possible world considered as actual, where the centre picks out the spatio-temporal
location of the speaker. This device models the contribution of indexicals in fixing reference. For more on this
see Chalmers 2004.
‘A’, which is transparently revealed to the concept user. An aspect of a property is itself a property. It follows that C affords its possessor a transparent concept of a property, namely A. To return to the above example of <being a sphere the same mass as the Earth>, the possessor of this concept transparently understands what it is for there to be a sphere, and thereby possesses a transparent concept of sphericity.

Thus, if phenomenal concepts are translucent, the anti-physicalist can try to run the Transparency Conceivability Argument not in terms of an arbitrary (actually referring) phenomenal concept <Q> itself, but in terms of the transparent concept involved in <Q>, call it <Q*>. For if the physical facts ground and hence necessitate all the facts, then they ground and hence necessitate the fact that Q* is instantiated. Just as the possibility of P without the instantiation of Q is inconsistent with physicalism, so is the possibility of P without the instantiation of Q*. We can put this modified version of the Transparency Conceivability Argument as follows:

The Specialised Transparency Conceivability Argument

Premise 1 – ‘P and nothing has Q*’ is conceivably true.

Premise 2 – (TCP) If a transparent sentence is conceivably true, then it’s possibly true.

Premise 3 – ‘P & nothing has Q*’ is transparent.

Conclusion – ‘P & nothing has Q*’ is possibly true, and so (pure) physicalism is false.

Whether or not premise 1 of the Specialised Transparency Conceivability Argument is true, however, will depend on the specific manner in which the translucency of phenomenal concepts is cashed out. It may perhaps be that phenomenal concepts reveal merely causal structural features of conscious states, causal structural features which may turn out to be instantiated in the physical structure human brains and hence implied by P. Physicalist Robert Schroer, for example, has an account of phenomenal concepts according to which they reveal the internal structure of phenomenal qualities, a view I will discuss in more detail in the next chapter. Suppose that Q is identical with some neurophysiological property a priori entailed by P, which has structure S as part of its essential nature. Suppose further that S is the aspect of Q revealed by <Q>, i.e. ‘S’ and ‘Q*’ are analytically equivalent. It
follows that ‘P and nothing has Q*’ is not conceivably true, as P involves a term which is analytically equivalent with Q*. In this case, the first premise of the Specialised Transparency Conceivability Argument is false and the argument fails.

For the specialised Transparency Conceivability Argument to work, the aspects of conscious states to which we have a priori access must be ones which are *not* a priori entailed by the physical facts, and hence ones which it is conceivable that zombies lack. Correspondingly if the physicalist wants to resist the Specialised Transparency Conceivability Argument (without denying TCP – we will consider that option later), then she must give an account of phenomenal concepts such that they are either opaque, or reveal only properties which are a priori entailed by the physical facts. Defending Phenomenal Translucency with this restriction is not straightforward.

We might usefully distinguish, then, between *Hard Phenomenal Translucency*, according to which phenomenal concepts reveal essential features of phenomenal qualities that are *not* a priori entailed by the physical facts, and *Soft Phenomenal Translucency*, according to which phenomenal concepts only reveal essential features of phenomenal qualities that are *a* priori entailed by the physical facts. If the physicalist can make sense of Soft Phenomenal Translucency, as Schroer tries to do, she can resist the Specialised Transparency Conceivability Argument.

In summary, TCP gives us a simple principle linking conceivability and possibility which (i) is not subject to Kripkean counterexample, (ii) does not beg the question against the physicalist by ruling out radically opaque or translucent concepts. 2D-CP is stronger than TCP, and if we can justify its adoption the refutation of physicalism is more straightforward: it is entailed by the Direct Reference Thesis. However, 2D-CP is supposedly justified by the alleged need to link conceivability to possibility; given that the weaker TCP also links conceivability to possibility we have no justification for going beyond TCP and adopting 2D-CP.

Framing the conceivability argument in terms of TCP rather than 2D-CP does allow the physicalist a means of resistance which doesn’t require breaking the link between
conceivability and possibility: she can defend *Radical Phenomenal Opacity* or *Soft Phenomenal Translucency*. In order to complete the argument the anti-physicalist must not only defend TCP itself, but must also defend *Phenomenal Transparency* or *Hard Phenomenal Translucency*. In the next chapter I will defend *Phenomenal Transparency*, before going on to consider the plausibility of TCP itself.