An Analysis and Defence of the Two-Dimensional Zombie Argument

Jiahao Wu

A thesis submitted in fulfilment of the requirements for the degree of Master of Philosophy

Department of Philosophy

Faculty of Arts and Social Sciences

The University of Sydney, 2023

Abstract

This thesis presents an analysis and defence of the two-dimensional zombie argument against physicalism by David Chalmers (2009). Put simply, the zombie argument uses the conceivability of zombies in an attempt to defeat physicalism, where zombies are physical duplicates of humans that lack consciousness and physicalism is the thesis that everything is either physical or supervenes on the physical. Despite the zombie argument being one of the most influential contemporary anti-physicalist arguments, the two-dimensional zombie argument, which is a refined version of the original zombie argument, remains relatively unknown among philosophers. In this thesis, I aim to clarify the two-dimensional zombie argument and defend it against four of the recent objections. The thesis is divided into four chapters. Chapter 1 is dedicated to introducing the two-dimensional zombie argument and elucidating the key concepts involved in the argument. Chapter 2 aims to provide an overall summary of the past discussions of the argument by mentioning the major objections made to the argument and defences against these objections. In Chapter 3, I provide detailed defences against objections mentioned in three of the more recent papers: one by Phillip Goff and David Papineau (2014); one by Daniel Stoljar (2020); and one by Eugen Fischer and Justin Sytsma (2021). In Chapter 4, I share my speculations on how language and intuitions might be the roots of many disputes over the argument and how further progress could be made. At the end, I conclude that the two-dimensional zombie argument, despite the large number of objections to it, remains in a highly defensible position. Once the argument is properly understood, there seems to be a lack of knockdown objections. At the same time, further progress can still be made by eliminating verbal misunderstandings and attempting to justify the intuitions involved.

Statement of Originality

This is to certify that to the best of my knowledge, the content of this thesis is my own work. This thesis has not been submitted for any degree or other purposes.

I certify that the intellectual content of this thesis is the product of my own work and that all the assistance received in preparing this thesis and sources have been acknowledged.

Jiahao Wu

Acknowledgements

First and foremost, I would like to thank the supervisory team. I am most indebted to David Braddon-Mitchell, my lead supervisor, who certainly has the biggest philosophical influence on me. Without his guidance, I would not have been able to complete this thesis. My associate supervisor, Peter Godfrey-Smith, has also played an important role in the development of my philosophical writing and reasoning. Michael Nielsen, despite only being on the team for the last few months of my candidature, has also provided some very helpful discussions and advice. It has been a pleasure and an honour working with them.

Special thanks to our postgraduate coordinator, Peter Anstey, who has been extremely helpful and responsive in dealing with my administrative problems during my candidature. His good work undoubtedly plays an important role in helping me complete my degree.

Last but not least, I would like to thank my family and my partner, who have provided constant support for my philosophical pursuits. I understand that not everyone has the opportunity to study what they are passionate about on a full-time basis, and I am grateful for the opportunity I have been given.

Table of Contents

Introduction

There are quite a few big questions in philosophy, and the mind-body problem is one of them. Put simply, the mind-body problem concerns the relationship between the mental and the physical. At the core of the mind-body problem lies the problem of consciousness. Consciousness is one of, if not the most, mysterious phenomena. We usually consider rocks and chairs as nonconscious, but dogs and humans as conscious, but why is that the case? If the universe once contained no conscious agent, how can consciousness be formed out of nonconscious matters? At what stage does consciousness first appear? Or has consciousness always been here? We, as humans, have always been fascinated by consciousness. At the same time, there are other strong motivations for us to develop a thorough understanding of consciousness. For example, we usually only assign moral status to conscious agents; no one cares about harming a piece of rock, but people debate about whether lobsters should be killed instantly before being thrown into boiling water to prevent suffering. Our knowledge of whether an agent is conscious has moral and other implications. With the rapid development of artificial intelligence over the past few decades, a consistent theory of consciousness is needed more than ever, as the question of whether artificial systems are capable of being conscious will only grow more and more relevant.

One solution to the problem of consciousness, and to the mind-body problem in general, is to adopt physicalism. Physicalism, in the most basic sense, is the metaphysical thesis that everything is physical. Physicalism sees consciousness, and any other mental state, to be identical to, or at least supervene on, some physical states. So, all we can know about consciousness will be revealed after we have learned enough about the physical world. Physicalism is quite a popular view adopted by many contemporary philosophers, and there are good reasons for that. First, the view is simple and elegant, as the physical serves as the sole fundamental metaphysical base of our world. But more importantly, physics is often considered as the most influential and fundamental of all sciences. All sciences are constrained by the laws of physics, and most natural phenomena appear to ultimately have some kind of physical explanation. So, it seems reasonable to suppose that consciousness might also one day obtain a physical explanation. In fact, plenty of research in neuroscience has been done on the subject of consciousness recently, and many neuroscientists believe that findings in neuroscience will ultimately reveal the nature of consciousness.

Despite being the mainstream view in philosophy of mind, not everyone considers physicalism as a satisfactory solution. Many well-known questions have been raised against physicalism, and many of them focus on how physicalism fails to properly explain

consciousness. For example, Nagel (1974) raises the well-known argument that no amount of physical information tells us about the conscious experiences a bat possesses, and Levine (1983) famously points out that physical explanations of consciousness always leave a significant explanatory gap.[1] Dualist intuitions of this sort are strong, even for some physicalists. So, to demonstrate that physicalism is true, physicalists must somehow explain away the dualist intuitions or at least show that their physicalist intuitions matter more than the dualist ones. Whether physicalism is true remains a topic of active heated debate in philosophy of mind.

One of the most influential contemporary arguments against physicalism is the zombie argument (also called "the conceivability argument"). The argument starts with the conceivability of zombies, creatures that are physically identical to humans but lack consciousness. Then, the argument uses the conceivability of zombies to argue for the possibility of zombies. After that, the argument points out the incompatibility between the possibility of zombies and physicalism, and by pointing out the incompatibility, the argument concludes that physicalism is false.

The idea of zombies is not novel, as similar ideas were previously mentioned by philosophers such as Campbell (1970), Kirk (1974), and Nagel (1974). However, the most discussed usage of the zombie idea arguably comes from Chalmers. Ever since the zombie argument was presented in Chalmers's (1996) book *The Conscious Mind: In Search of a Fundamental Theory*, the argument has sparked plenty of discussions. Many people were at least sympathetic to the overall idea of the argument, but many objections were also raised against it. Most objections either focus on rejecting the conceivability of zombies (e.g., Thomas, 1998; Shoemaker, 1999; Balog, 1999; Clark, 2000; Perry, 2001; Worley, 2003; Bailey, 2007; Kirk, 2005, 2008), or rejecting the link between conceivability and possibility (e.g., Shoemaker, 1999; Yablo, 1999; Loar, 1999; Hill & McLaughlin, 1999; Papineau, 2002), but some objections also focus on other aspects of the argument (e.g., Stoljar, 2001b; Braddon-Mitchell, 2003; Frankish, 2007; Piccinini, 2017). Shortly after the publication of his influential 1996 book, Chalmers himself wrote several articles to further clarify his argument and defend it against some of the proposed objections (1999, 2002, 2003a, 2003b, 2004a, 2006a, 2009). The most notable work is arguably his 2009 article *The Two-Dimensional Argument against Materialism*. In this paper, he provides a more elaborate version of the zombie argument, which

[1] It should be noted that neither Nagel nor Levine are considered as dualists. Nonetheless, the two papers they wrote demonstrate the strong dualist intuitions physicalism must overcome.

I shall call the "two-dimensional zombie argument," by explicitly infusing two-dimensional semantics into the argument. In this paper, besides presenting the two-dimensional zombie argument, he also provides a thorough analysis of the potential objections and how they might be addressed.

Similar to Chalmers and many other dualists, I share strong dualist intuitions. Even though I admit that physics is a powerful tool in explaining our world and that physicalism presents a very elegant picture of our world, I simply do not see how consciousness can be explained by the physical. I think that physicalism is false, and to defeat physicalism, the two-dimensional zombie argument seems to be the perfect weapon. I believe that, when the two-dimensional zombie argument is understood properly, the argument is sound and capable of bringing down physicalism, or at least the kind of physicalism most people discuss in philosophy of mind. So, the current thesis uses Chalmers's 2009 paper as the main inspiration and attempts to provide a detailed analysis and further defence of the two-dimensional zombie argument.

The main goal of this thesis is a humble one, which is to further defend the argument from four of the more recent objections. But to do so, I will first need to clarify the argument by analysing the concepts and ideas involved. I will also summarise the work that has been done on the argument over the past few decades by introducing the different objections made to the argument and how they have been replied to. Near the end of the thesis, I will share some of my own meta-philosophical speculations on why so many disputes exist over the argument and how they might be resolved. Even though I would love to settle the debate once and for all, I do not have such ambitions with this thesis. Plenty of very smart people have contributed to the debate over the past few decades, but still no consensus has been reached. Here, I only hope to shed a bit more light on the debate and make the argument seem a bit more credible.

So here is the trajectory of the thesis. Chapter 1 aims to provide an introduction and clarification of the two-dimensional zombie argument. Most philosophers are at least acquainted with the zombie argument, but few are familiar with all the details and subtleties involved in the argument, especially those in its two-dimensional version. So, after briefly introducing the argument at the beginning of the chapter, I will go through all of the core ideas and topics involved in the argument. These ideas and topics include zombies, two-dimensional semantics, possibility, supervenience, conceivability, consciousness, and physicalism. Understanding these ideas is crucial for understanding the argument, as meaningless disputes can easily be generated when these ideas are not properly understood. In this chapter of the thesis, I will discuss how these ideas are usually interpreted in philosophy and, in contrast, how

they should be interpreted in the context of the argument. After clarifying all the ideas, I will pull my focus back to the argument and clarify all the premises and the conclusion. Overall, Chapter 1 intends to provide the background knowledge and specify the exact argument that we are dealing with in this thesis.

The aim of Chapter 2 is to provide a summary of the recent work that has been done on the argument. Since the argument is clearly valid, the soundness of the argument depends on whether all the premises are true. This chapter is divided into four sections, with each section dealing with one of the four premises. In each section, I will first provide the reasons for supporting the premise. Some ideas here will come from Chalmers's previous work, and some will come from my own reasoning. Then, I will summarise the major objections proposed to the given premise. Due to the sheer number of objections that have been raised to the argument, my summaries here might not be exhaustive, but I will at least try to include all the well-known and influential ones. I will also mention the defences that have been made against these objections. Overall, Chapter 2 outlines the previous attempts at refuting and defending the argument, which acts as a useful background for my own attempts to defend the argument in the next chapter.

Chapter 3 will be the place where I attempt to make my main contribution to the debate. In this chapter, I will be dealing with three papers that have proposed objections to the argument: one from Goff and Papineau (2014), one from Stoljar (2020), and one from Fischer and Sytsma (2021). In the first paper, Goff and Papineau both raise their own individual arguments for a kind of modal dualism, which opens up the possibility for strong necessities and breaks the link between conceivability and possibility. In the second paper, Stoljar points out an inconsistency between Chalmers's phenomenal functionalism and his zombie argument, and therefore argues that one of them must be given up. In the third paper, Fischer and Sytsma use empirical data to demonstrate that many people fail to find zombies conceivable, especially when linguistic salience bias is taken into account. The three objections are selected because they are all relatively recent and interesting enough to discuss. The recency of these papers also means that these objections are not yet thoroughly explored. Even though these objections have not yet attracted a lot of attention, the authors of the first two objections are already well-known for their contributions towards the mind-body problem, and the third objection is unique in the sense that it relies on experimental data to refute the argument. Since there are three papers to discuss in this chapter, the chapter will be divided into three sections, with one section for each paper. In each section, I will first introduce the objection involved and the pre-existing replies to it, if there is any. Then, I will provide my own analysis and explain why each objection fails.

In Chapter 4, I will share my speculations on the nature of the debate over the zombie argument. This is the place where I will dip my toes into meta-philosophy and philosophy of language, and deal with questions such as what the roots of disputes in philosophy are, how these disputes can be resolved, and how progress can be made on philosophical questions. These questions might not seem to be directly linked to the zombie argument, but they are questions we should answer before we can settle all the disputes over the zombie argument once and for all. This chapter will be divided into two sections, with one focusing on language and the other focusing on intuitions. In the first section, I will first introduce a framework of language and concepts that I think serves well for philosophical discussions. I will then discuss how verbal disputes might be generated and how they can be dealt with. Afterwards, I will discuss some of the verbal disputes that exist in the debate over the zombie argument. In the second section, I will first introduce the concept of intuition that I am interested in here. Then, I will discuss how these intuitions matter in philosophy and what they imply about philosophical progress. Finally, I will discuss how some disputes over the zombie argument might be related to our differences in intuitions, and what we might be able to do about these differences. Overall, this chapter intends to analyse the zombie argument from a broader perspective and speculate about how progress could be made on the debate over the zombie argument.

Overall, as a thesis, the first two chapters aim to demonstrate my understanding of the subject matter and provide the necessary background for my later discussions, while the last two chapters aim to demonstrate my own reasoning and my contributions to the debate. Although the current thesis does not have the ambition of completely resolving the debate over the zombie argument, it at least intends to make a modest contribution by clarifying the argument and ruling out a few objections. There is no doubt that the problem of consciousness, and the mind-body problem in general, is extremely perplexing, but I am optimistic that progress can be made on them, and we can start by resolving the debate over the zombie argument.

Chapter 1: The Two-Dimensional Zombie Argument

In this chapter, I will be introducing and clarifying the subject matter of the thesis – the two-dimensional zombie argument. I will start with the basic version of the zombie argument that most philosophers are familiar with. Then, I will go through all the key concepts involved in the two-dimensional argument. Last but not least, I will utilise the discussions on these key concepts to clarify the two-dimensional argument by going through each premise and the conclusion.

1.1 Introducing the Argument

The zombie argument, sometimes also known as the conceivability argument, is one of the most influential arguments against physicalism in contemporary philosophy. The basic version of the argument runs as follows:

(P_b1) Zombies are conceivable.

(P_b2) If zombies are conceivable, zombies are possible.

(P_b3) Physicalism is incompatible with the possibility of zombies.

(C_b) Physicalism is false.

Zombies, put simply, are physical duplicates of humans that lack consciousness. The argument is clearly valid: P_b1 gives us the conceivability of zombies; the conditional in P_b2 combined with the conceivability of zombies gives us the possibility of zombies; and the conditional in P_b3 combined with the possibility of zombies gives us the falsity of physicalism, which is the conclusion. The soundness of the argument, on the other hand, depends on the truths of the premises. To evaluate this argument, we will need to understand all the concepts involved and the relations between these concepts. Many questions need to be answered: What counts as being physical? What is consciousness? What is conceivability? What sort of modality is in discussion here? What is the link between conceivability and possibility? What is physicalism? Why is physicalism incompatible with the possibility of zombies?

Before we get into answering these questions, it is important to realise that although the basic version of the argument is the most well-known, it is not the most sophisticated. Lots of ambiguities exist in the basic version of the argument. David Chalmers (2009) later proposed a more refined version of the argument by infusing the argument with two-dimensional semantics. I shall call it the "two-dimensional zombie argument" here, and it goes like this:

(P1) Zombies are ideally primarily conceivable.

(P2) If zombies are ideally primarily conceivable, zombies are primarily possible.

(P3) If zombies are primarily possible, zombies are either secondarily possible or Russellian monism is true.

(P4) If zombies are secondarily possible, physicalism is false.

(C) Physicalism is false or Russellian monism is true.

It is the two-dimensional zombie argument that I want to defend in this thesis. When I mention the zombie argument in this thesis, I will be referring to the two-dimensional version by default. This formulation of the argument also gives rise to some further questions: What do the terms "primary" and "secondary" mean here? What is the difference between ideal conceivability and non-ideal conceivability? What is Russellian monism? These questions, including the ones mentioned earlier, will be answered in this chapter. In the following, I will explain and analyse the core concepts involved in the argument. After all the concepts are sufficiently discussed, I will come back to the argument in Chapter 1.7 and clarify the premises and conclusion.

1.2 Zombies

Philosophical zombies (or "zombies" for short), as mentioned earlier, are physical duplicates of humans that lack consciousness. The natural questions to ask next are what counts as being physical and what counts as being conscious. I will leave the discussions on these questions for later, as both of them deserve a more detailed analysis. For now, I shall just move on with the intuitive senses of being physical and being conscious.

The term "zombie" is most often used to refer to movie zombies by laypeople. Movie zombies are not physically identical to humans; they often appear and behave differently, which allows us to easily distinguish them from humans via their behaviours and appearances. The specifics of movie zombies differ from movie to movie, but the common features of movie zombies include aggressiveness or slow movements, being incapable of communicating with humans, pale skin, a rotting body, and eating brains or flesh of humans.

Philosophical zombies do not share any of the above features. Since philosophical zombies are physically identical to humans, they also share identical appearances and behaviours.[2] To create a zombie, we can take a conscious human H_c and take away its consciousness while leaving it physically intact. The result will be a philosophical zombie H_z. We shall call H_z the zombie twin of H_c as they share the same physical properties. H_z and H_c will behave and appear exactly the same, with the only difference being that H_c has conscious

[2] There is an assumption here that behaviours and appearances supervene on the physical, but this assumption should be well accepted.

experiences but H_Z does not. There is no way to tell them apart from the third-person perspective.

Even though philosophical zombies have no conscious experience, some of them might seem to be discussing consciousness. Suppose we create a zombie twin of David Chalmers that shares all the behaviours and appearances of the real David Chalmers. This zombie twin will be writing books on consciousness and talking about consciousness, at least in the superficial sense. This zombie will even claim that he is conscious, despite the fact that he is not. This is called "the paradox of phenomenal judgements," and it certainly seems counterintuitive to many people. Proponents of the zombie argument, sometimes called the "zombists,"[3] usually argue that although cases like this seem counterintuitive at first, there is no underlying incoherence. Still, some philosophers think that this paradox is problematic enough for us to reject the possibility of zombies (see Thomas, 1998; Perry, 2001). To fully understand this case, I think, we need to examine the theory of reference and the theory of content that we have. As a zombist myself, I think it is helpful to simply think about a machine that displays the word "consciousness" or pronounces the word "consciousness" with its speakers. Most people do not think such a machine has consciousness. Its consciousness-related behaviours can be fully explained by a complete physical analysis of the system. Once we learn about all the physical mechanisms of this system, there is no more mystery left about its consciousness-related behaviours. The same idea can be applied to humans. This is by no means a detailed answer to the paradox, but I do think it provides some intuitive appeal to why the paradox might not be as problematic as it seems. More on phenomenal judgements will be discussed in Chapter 2.

Although most of the contemporary discussions on zombies and the zombie argument can be traced back to David Chalmers's (1996) *The Conscious Mind*, the idea of zombies has been entertained by many philosophers before that. Campbell (1970) mentioned the idea of an "imitation man" that duplicates people's behaviours but possesses no phenomenal properties. Kirk (1974) considered a similar idea where an organism is "indistinguishable from a normal human being in all anatomical, behavioural, and other observable respects, yet insentient" (p. 43). Nagel (1974) also described the idea of "robots or automata that behaved like people though they experienced nothing" (p. 436). Philosophers like Block (1980a, 1980b) and Shoemaker (1975) were later engaged in heated discussions about the absent qualia idea where two functionally identical systems are qualitatively different. In short, the zombie intuition has

[3] The term "zombist" is most notably used by Keith Frankish (2007).

always been prevalent in philosophy of mind, even though some philosophers do not think zombies are possible.

Most philosophers, including the zombists, think that zombies do not exist in the actual world. They are mere hypothetical creatures. At the same time, even if zombies actually existed among us, there would be no way of knowing, as we only have direct access to our own consciousness from the first-person perspective and not to other humans' consciousness from the third-person perspective.[4] Nonetheless, discussions on zombies mostly focus on the possibility of zombies. To defeat physicalism, we do not need zombies to actually exist, but only their possibility. The reason lies in the formulation of physicalism, which will be discussed later.

Besides the idea of zombies, discussions on the zombie argument also often involve the idea of zombie worlds. Usually, a zombie world is the bare physical duplicate of our world that lacks any conscious agent or conscious phenomenon. However, we can also construct a world where conscious agents and zombies co-exist. I shall call the former the "complete zombie world" and the latter the "partial zombie world." It should be noted that being the bare physical duplicate of our world alone does not make a zombie world. Physicalism claims that our world is purely physical. So, according to physicalism, if consciousness does exist, duplicating the physical features of our world will automatically give us consciousness. So, besides being physically identical to our world, a zombie world must contain zombies that are not conscious.

1.3 Two-Dimensional Semantics

Before I get to the topics of modality, consciousness, and physicalism, I need to first introduce two-dimensional semantics. The reason will become obvious later. Two-dimensional semantics is essential for us to understand the distinction between primary conceivability and secondary conceivability, as well as the distinction between primary possibility and secondary possibility. Furthermore, the two-dimensional zombie argument analyses the concept of *physical* and the concept of *consciousness* based on two-dimensional semantics. Introducing two-dimensional semantics now will lay the foundation for later discussions. In this section, I will first go through the relationship between language and concepts. This will be followed by a brief introduction to the intensional theory of semantics, and last but not least, I will introduce two-dimensional semantics.

---

[4] Physicalists who consider consciousness to be *a priori* entailed by physical states might disagree with this. For them, knowing whether a system is conscious is just knowing specific physical information about the system. Therefore, they will find the idea of zombies incoherent.

1.3.1 Language and Concepts

Language is the main tool people use to communicate, both in everyday life and in academic discussions. But very often, language itself is not the subject of interest. What we care about are often the ideas the sentences or words represent. However, the link between language and the idea is not always straightforward. If such a link is not handled properly, language can become a major source of confusion. For example, confusion can arise when two parties attach different meanings to the same word without knowing it or when two parties use different words to mean the same thing. One example might be two people debating whether their friend likes orange, but in fact, their friend likes eating oranges and dislikes the orange colour. Both people might be right in this apparent debate, and the dispute is only superficial. In our case, the term "orange" itself is not capable of identifying the meaning that we like to communicate. Similarly, we have seen earlier that the term "zombie" might identify different meanings. Someone might argue for the possibility of zombies by explaining the lack of incoherence in the idea of virus-infected rotting creatures that crave human flesh. It should be obvious why such an argument does not work. Clarifying the meanings of potentially ambiguous terms should be the priority of any philosophical discussion.

The solution I adopt here is to focus on concepts instead of words. In the scenario presented earlier, we can say that the two people had two different concepts for the term "orange." Concepts here act as the carriers of meaning. There can be multiple concepts associated with one term, or one concept associated with multiple terms.

There is another problem. How do we know what concepts people possess? We know that the two people in our scenario had two distinct concepts for the term "orange," but what are these concepts? My suggestion here is to use descriptions. For example, the concept of *bachelor* is *man who has never been married*, where the description "man who has never been married" acts as a clarification of what the concept represented by the term "bachelor" is. Coming up with an accurate description requires a thorough conceptual analysis (Jackson, 1998). The process of coming up with descriptions for concepts closely resembles the process of coming up with definitions for terms. The use of definition has indeed received some well-known criticism over the past few decades, such as from Quine (1951) and Kripke (1972). I will deal with these more complicated issues with language and concepts in the last chapter of the thesis. For now, I shall just take the framework introduced above for granted, where meanings are carried by concepts and the rough content of concepts can be communicated via descriptions.

1.3.2 Intension and Extension

What is meaning? We know that concepts pick out certain things in our world, which we call the referents of the concept. The set of referents constitutes the extension of the concept. For example, the concept of *rabbit* is used to refer to actual rabbits in our world, and as a result, these rabbits are picked out by the concept of *rabbit*. The actual rabbits are the referents of the concept, and the set of referents counts as the extension of the concept. However, extensions do not seem to be the sole component of concepts. Two concepts might share the same extension without sharing the same meaning. To use Quine's (1951) and Putnam's (1975) example, the concept of *creature with a kidney* and the concept of *creature with a heart* perhaps share the same extension, but nonetheless, they differ in their meanings in some way. To explain this difference, we need intensions.

We can usually understand intension as an *a priori* function with possible worlds as inputs and extensions as outputs.[5] This function *f*, together with a possible world *w* as input, determines an extension *f (w)* as output. With possible world w_1, the extension will be $f(w_1)$; with possible world w_2, the extension will be $f(w_2)$, etc.[6] To utilise this idea with Putnam's example above, we can create one possible world where the concept of *creature with a kidney* and the concept of *creature with a heart* pick out two distinct extensions. If there is indeed such a possible world, it will show that the two concepts are associated with two distinct functions, which means the two concepts have distinct intensions. Intension is commonly considered as a better candidate for meaning than extension.

1.3.3 Primary Intension and Secondary Intension

Two-dimensional semantics further distinguishes between two types of intensions – *primary intension* and *secondary intension*,[7] where the primary intension picks out the extension from any possible world when the possible world is considered as actual, and the secondary intension picks out the extension from any possible world when the possible world

[5] This is associated with the idea that conceptual truths, such as "all bachelors are men who have never been married," are purely *a priori*. The truth of such a claim has nothing to do with any empirical fact. For more information, see Jackson (1998) and Chalmers (2012).

[6] This explanation of intension here relies on a sort of possible worlds semantics, which is not the only theory of semantics available. However, discussions on other theories of semantics will be out of scope here.

[7] Frank Jackson (1998), another advocate of two-dimensional semantics, uses the term "A-intension" for primary intension and "C-intension" for secondary intension. For the sake of consistency, I will stick with Chalmers's choice of words.

is considered as counterfactual (Chalmers, 2004b, 2006b, 2006c). To better explain this idea, let me first talk about *a posteriori* necessities proposed by Kripke (1972).

It was traditionally considered that only *a priori* claims can be necessary. This means claims such as "water is H_2O" can only be contingent, as the truth of the claim depends on *a posteriori* information. Kripke thinks otherwise. Consider the Twin Earth thought experiment by Putnam (1975). Twin Earth is a place that is almost identical to Earth, except that there is no H_2O but only XYZ.[8] However, XYZ shares the same apparent watery properties as H_2O such as being transparent, odourless, and drinkable. Furthermore, people from Twin Earth possess the identical concept of *water* as people from Earth, which is *the actual local watery stuff*.[9] It seems to make sense that when the concept of *water* is used by people on Twin Earth, the concept picks out XYZ, as XYZ is the local watery stuff for them, not H_2O. However, Kripke argues that the concept of *water* always refers to H_2O, since only H_2O is the actual local watery stuff. The concept of *water* here is considered a rigid designator, whose referents are fixed by how the actual world turns out to be. For Kripke, identity claims are always necessary. As residents on Earth, the concept of *water* is fixed to only refer to H_2O due to H_2O being the actual local watery stuff. When we think about Twin Earth, our concept of *water* does not pick out the XYZ there, as XYZ is not the actual local watery stuff. Therefore, the claim "water is H_2O" is true on Twin Earth, and it is necessarily true across all possible worlds, which makes it an *a posteriori* necessity.

The idea of *a posteriori* necessity (also sometimes called "Kripkean necessity") has been well accepted in contemporary philosophy. It is also sometimes used as a weapon to attack the zombie argument, which I will talk about later. However, it also seems like *a posteriori* necessity only captures some, but not all, of our intuitions. There is no apparent contingency in claims such as "water is water," but an apparent contingency in claims such as "water is H_2O" and "Hesperus is Phosphorus." If the actual world had turned out to be another way, where the local watery stuff was found to be XYZ, we would have considered water to be XYZ. Two-dimensional semantics provides an answer for this apparent contingency.

The key idea of two-dimensional semantics is the two ways of considering a world: as actual and as counterfactual. The primary intension is involved when a world is considered as

[8] Here, I assume that Twin Earth is not a place in our universe but a place in another possible world. Therefore, we can say that Twin Earth does not exist actually.

[9] Once again, the concept of *water* here is simplified. A more accurate concept might be *the dominant essence that exists in the actual local stuff that possesses watery properties such as being transparent, odourless, and drinkable*. However, the concept provided in the main text is sufficient for demonstrating the relevant ideas.

actual, and the secondary intension is involved when a world is considered as counterfactual. Take the case of water earlier. There seems to be no incoherence in the existence of Twin Earth where the local watery stuff is XYZ, not H_2O. I think most of us can agree with the possibility of such a world. The reason Kripke thinks that "water is H_2O" is still true on Twin Earth is because he thinks that the local watery stuff on Twin Earth does not count as the referent of our concept of *water*. Our concept of *water* always picks out H_2O as the referent because H_2O is the actual local watery stuff. Earth, but not Twin Earth, exists in the actual world.[10] So, according to Kripke, the apparent contingency in "water is H_2O" is due to our mistake of thinking that our concept of *water* refers to the non-actual local watery stuff.

However, if different possible worlds are allowed to be considered as actual, a different story could be told. Suppose the possible world where Twin Earth exists turns out to be actual; the concept of *water* will then refer to XYZs since the concept is *the actual local watery stuff*. According to two-dimensional semantics, concepts such as "water" and "Hesperus," similar to concepts such as "I" and "now," are indexical. What they refer to depends on which world counts as the actual world or which context counts as the relevant context. So, the primary intension of the concept of *water* picks out the XYZs on Twin Earth, while the secondary intension picks out the H_2O on Twin Earth.[11] Two-dimensional semantics explains the apparent contingency in "water is H_2O" by saying that the claim is necessary when our world is fixed as actual, but the claim is merely contingent when other possible worlds can also be considered as actual.

What two-dimensional semantics says is that there are actually two kinds of necessity. The first kind is based on the primary intension of the concept, and the second kind, the kind that Kripke (1972) mentions, is based on the secondary intension. The first kind is what Gareth Evans (1979) calls "deep necessity," and the second kind is what Chalmers (1996) calls "superficial necessity." According to Chalmers, statements like "water is H2O," if they are necessary at all, can only be superficial necessities but not deep necessities.

---

[10] For the sake of simplicity, I am considering Earth and Twin Earth to exist in two different worlds. It is certainly fine to run the story with both Earth and Twin Earth in the same world. If that is the case, we might say that the secondary intension of water refers to H_2O on both Earth and Twin Earth because Earth is local and Twin Earth is not. Two-dimensional semantics is not only capable of explaining our modal intuitions but also capable of explaining our intuitions about the meanings of indexical words in different contexts.

[11] There is no H2O on Twin Earth, which means the extension picked out by the secondary intension of the concept of *water* will be empty.

According to two-dimensional semantics, each concept has two intensions. However, the two intensions can be identical. A concept with identical primary and secondary intensions will pick out the same class of properties and things in every possible world, regardless of whether the possible worlds are considered as actual or counterfactual. Using Philip Goff's (2011) terminology, we might consider these concepts as transparent, and concepts with distinct primary and secondary intensions as opaque.[12] For an identity claim to be an *a posteriori* necessary, at least one of the concepts involved has to be opaque. The distinction between different kinds of concepts will become more prominent when we start analysing the concept of *consciousness* and the concept of *physical* later.

1.4 Possibility, Supervenience, and Conceivability

Possibility and conceivability lie at the heart of the zombie argument. Most objections to the zombie argument focus on either denying the conceivability of zombies or the link between conceivability and possibility. The way we understand possibility and conceivability has a huge influence on how we evaluate the zombie argument.

Another important idea involved in the argument is the idea of supervenience. The possibility of zombies is capable of undermining physicalism because it shows that consciousness fails to supervene on the physical. Possibility, supervenience and conceivability are closely related ideas. Understanding these ideas will be an essential part of understanding the zombie argument. In this section, I will focus on clarifying these ideas and how they are understood under the two-dimensional semantics framework.

1.4.1 Possibility

The world we live in is the actual world. Our world has its own way. However, there is a sense that our world could have turned out to be otherwise. For example, David Chalmers could have become a mathematician instead of a philosopher, or I could have decided to write a thesis on ethics instead of metaphysics, but one plus one could never equal three. It seems to us that certain things could have happened but others could not. Theories of modality aim to explain our modal intuitions.

[12] When Philip Goff proposed the distinction between transparent, translucent, mildly opaque, and radically opaque concepts, he did not specifically have two-dimensional semantics in mind. However, his framework can be explained using two-dimensional semantics quite nicely, except for the idea of radically opaque concepts. For more information on this distinction, see Chapter 3.1.

1.4.1.1 Possible Worlds and Three Kinds of Possibility. Discussions on modality often rely on the idea of possible worlds. Let us say that P is a proposition. P is possible if and only if P is true in some of the possible worlds; P is necessary if and only if P is true in all possible worlds (which includes the actual world); and P is impossible if and only if P is true in no possible world. For example, the statement "there are flying pigs" is possible if and only if flying pigs exist in some possible worlds, and the statement "1+1=3" is impossible if and only if such a statement is false in every single possible world.[13]

Traditionally, there are three kinds of modalities commonly discussed in metaphysics: logical modality, metaphysical modality, and nomic (natural) modality. It is often considered that each kind of modality is accompanied by its own space of possible worlds. A world is logically possible if and only if nothing in the world contradicts the laws of logic, and a similar idea can be applied to the metaphysical and nomic modality, but with the laws of metaphysics and the laws of nature accordingly. The space of the logically possible worlds is the widest, which means some worlds might be logically possible without being metaphysically or naturally possible, but any world that is metaphysically or nomically possible must also be logically possible. In the following, however, I will introduce and support modal monism, a view that considers there to only be one primitive modal space. If modal monism is true, there is no need to propose a distinct space of metaphysically possible worlds, and that metaphysical possibility either makes no sense or can be explained using only the space of logically possible worlds and two-dimensional semantics.

1.4.1.2 Modality under Two-Dimensional Semantics. Instead of focusing on the distinction between logical, metaphysical, and nomic modality, two-dimensional semantics focuses on the distinction between primary and secondary modality (Chalmers, 2002, 2009). The primary/secondary distinction in modality will also be the one that I focus on in this thesis. Let us use proposition P as an example again. P is primarily possible (1-possible) if and only if its primary intension is true in some possible worlds, or we can say that P is primarily possible if and only if P is true in some possible worlds considered as actual. On the other hand, P is secondarily possible if and only if P is true in some possible worlds considered as counterfactual. To use the water example, "water is H_2O" is primarily possible but secondarily necessary, since the primary intension of the claim is false in some possible worlds but the secondary intension is true in all possible worlds. The primary modality is sometimes called

[13] More accurately, it is the proposition represented by the statement that can be possible or impossible.

the "epistemic modality," and the secondary modality is sometimes called the "subjunctive modality" (Chalmers, 2004).

What is the relationship between the primary/secondary distinction and the logical/metaphysical/nomic distinction? First, we can start with the space of logically possible worlds, which is the widest space that we can get. These are the worlds that are logically coherent. We can then generate the space of nomically possible worlds by only including the worlds that share the same laws of nature as our world. Nomic modality is based on the space of nomically possible worlds, but since nomic modality is not the focus of our discussion, I will not get into the details here. I will instead focus on the distinction between logical modality and metaphysical modality, as such distinction is central to our discussion.

The two-dimensional picture Chalmers (2002, 2009) sketches considers there to be only one primitive space of possible worlds – the space of logically possible worlds. This view is called "modal monism." The distinct feature of modal monism is the lack of the space of metaphysically possible worlds. But then, how do the two modalities differ if the same space of possible worlds is involved? The key, once again, lies in the two ways of considering a world. Consider the claim "water is XYZ" again. Many philosophers consider this claim as logically possible but metaphysically impossible, at least when using the Kripkean *a posteriori* necessity as the paradigm for metaphysical modality. Modal dualism, the view where two primitive spaces of possible worlds exist, seems to suggest that the claim is true at a possible world W that is logically possible but metaphysically impossible. According to modal dualism, the space of logically possible worlds contains W, but the space of metaphysically possible worlds does not. However, that is the case only if we do not adopt the two ways of thinking about possible worlds.

According to modal monism, it is not the case that W is included in the space of logically possible worlds but not in the space of metaphysically possible worlds. There is only one space of possible worlds, which contains W. The claim "water is XYZ" is true in W when W is considered as actual, but false in W when W is considered as counterfactual. There is no question concerning the nature of W; the only question here is whether the local watery stuff (XYZ) should count as water. According to two-dimensional semantics, the local watery stuff in W counts as water when W is considered as actual but does not count as water when W is considered as counterfactual. Two-dimensional semantics has no trouble accommodating the Kripkean cases, and there is no need for the space of metaphysically possible worlds.[14]

[14] Braddon-Mitchell and Jackson also argue a similar point in Section 4 of their 1996 book.

Some philosophers might still insist that there is a different kind of metaphysical modality that is different from Kripkean necessity and cannot be explained using one primitive space of possible worlds. If that is the case, a distinct space of metaphysically possible worlds might be needed. However, so far, there is little evidence or reason for what this kind of metaphysical modality might be. Proposing a distinct space of metaphysically possible worlds will require an explanation of what the laws of metaphysics are, which I do not think could be done easily. Chalmers (2002, 2009) himself also argues that such a space of metaphysically possible worlds appears to do no work and explain nothing. I totally agree with Chalmers here, and the only viable interpretation of metaphysical modality, in my opinion, is to interpret it as the secondary modality. In summary, Chalmers's two-dimensional semantics and modal monism demonstrate that one single primitive space of possible worlds plus two ways of considering each world give us all the tools we need to explain our intuitions on both logical and metaphysical modalities.

1.4.2 Supervenience

Supervenience is essential to physicalism. But what is supervenience? When we think of the objects, events, and facts in this world, we do not think of them as separate and unrelated. Instead, we think of these things as interconnected; there is a notion of dependence (for discussion, see Kim, 1984). Some of the dependency relationships can be captured by the concept of *supervenience*. Supervenience concerns relationships between sets of properties. The relationship can be described roughly as follows:

> A-properties supervene on B-properties if and only if there are no two objects (or possible scenarios) that are identical in terms of their B-properties but differ in terms of their A-properties.

To put it simply, A-properties are fixed by B-properties. In this case, we call the A-properties "supervenient properties" and the B-properties "base properties." Supervenience is also closely related to modality. If A-properties supervene on B-properties, there are no two possible objects/scenarios/worlds that differ in the distributions or characteristics of A-properties without also differing in the distributions of B-properties.

There is a distinction between local supervenience and global supervenience. A-properties supervene locally on B-properties when the A-properties of an individual or object are fixed by its B-properties. On the other hand, A-properties supervene globally on B-properties when the A-properties of a world are fixed by its B-properties. In short, local supervenience concerns individuals, and global supervenience concerns worlds. Local

supervenience is more difficult to achieve than global supervenience, which means there are situations where there is a global supervenience but not local supervenience between two sets of properties. This can happen sometimes when the supervenient properties are context-dependent.[15] In the discussion on physicalism, people mainly focus on global supervenience.

Besides the distinction between local and global supervenience, there is also the distinction between logical and nomic supervenience. A-properties supervene logically on B-properties when there are no two logically possible worlds that differ in A-properties without also differing in B-properties. On the other hand, A-properties supervene nomically on B-properties when there are no two nomically possible worlds that differ in A-properties without also differing in B-properties.

Metaphysical supervenience is more interesting. The kind of supervenience physicalism cares about is metaphysical supervenience, but as I have explained earlier, there is no distinct space of metaphysically possible worlds according to Chalmers's modal monism. The difference between logical supervenience and metaphysical supervenience then does not come from the space of possible worlds but from the way we look at those worlds. To use the example of water again. The distribution of water metaphysically supervenes on the distribution of H_2O, but it does not logically supervene. When the secondary intension of *water* is in play, there are no two possible worlds where the distributions of water differ without the distributions of H_2O also differing. But when the primary intension is in play, we do can find two possible worlds that differ in their distributions of water but do not differ in their distributions of H_2O. More specifically, we can conceive of two worlds where H_2O are sparsely distributed around both worlds, but one of the worlds also has an abundant amount of XYZ and the other does not. Both worlds have the same distribution of H_2O, but the H2O is only identified as water in the world that lacks XYZ. Therefore, under the two-dimensional semantics framework, metaphysical supervenience can be regarded as a variety of logical supervenience (Chalmers, 1996).

1.4.3 Conceivability

The zombie argument starts with the conceivability of zombies. If zombies are inconceivable, the argument fails at step one. Even if zombies are conceivable, the argument still fails if conceivability does not lead to possibility. What is conceivability? What is the

[15] One example of the failure of local supervenience used by Chalmers (1996) is the supervenience of biological properties on physical properties.

relationship between conceivability and possibility? These are the two core questions we need to deal with here.

It is not an easy task to explain what conceivability is. This is mostly because the concept of *conceivability* itself is quite ambiguous. When used in everyday English, the term "conceive" usually means *to form an idea or thought about something*. As Yablo (1993) points out, the act of conceiving almost always involves the appearance of possibility. Still, it is unclear what conceivability really is, and whether it is a guide to possibility.

As one of the major proponents of the zombie argument, David Chalmers has provided a framework for understanding conceivability himself. Both core questions are more or less answered by his framework. In the following, I will introduce Chalmers's account of conceivability by focusing on his (2002) paper *Does Conceivability Entail Possibility?*.

1.4.3.1 Three Dimensions of Conceivability by Chalmers. Chalmers identifies three dimensions of conceivability: negative/positive, ideal/*prima facie*, and primary/secondary. With the three dimensions, we can create 8 (2^3) different kinds of conceivability. Let me start with the negative/positive dimension first. The negative/positive dimension concerns how conceivability should be characterized. According to Chalmers, A statement S is negatively conceivable when S is not ruled out *a priori*. I also understand it as saying S is negatively conceivable when there is no incoherence in S, since it seems like S will be ruled out *a priori* if and only if there is any incoherence in S.

Positive conceivability, on the other hand, is a bit more complicated. Roughly speaking, we can say that S is positively conceivable when we can coherently imagine a situation where S is true. One problem with positive conceivability concerns the concept of *imagination*. Just like conceivability, "imagination" is not well-defined. We all have some understanding of what it is like to imagine something, but our imagination has a lot of limitations. For example, can we coherently imagine a purple apple? There seems to be little trouble doing that. However, when we attempt to imagine the purple apple, we do not seem to be able to depict all the details of the apple. We simply imagine the paradigmatic purple colour being thrown on a paradigmatic apple-shaped object. There are more problematic examples such as imagining a one-thousand-sided object and imagining a city with exactly 263 people. Both scenarios seem to be perfectly coherent, but I do not think anyone is capable of vividly imagining either scenario in their mind. For this reason, I prefer dealing with negative conceivability rather than positive conceivability. Nonetheless, the concept of *positive conceivability* seems to resemble what philosophers mean by the term "conceivability" more closely. It should also be noted that positive conceivability entails negative conceivability, but the reverse is not the case.

The second dimension concerns the distinction between ideal and *prima facie* conceivability. The phrase "*prima facie*" means *at first sight*. We call something *prima facie* when it is based on first impressions. So, something is *prima facie* conceivable when it is conceivable at first sight, without having undergone thorough reasoning. Ideal conceivability, on the other hand, requires perfect or ideal reasoning. Something is ideally conceivable if its conceivability stands under ideal rational reasoning. A question arises: What is ideal reasoning? This is certainly an important question in philosophy, but not a question that we need to solve here. Chalmers (2002) notices that the idea of ideal reasoning is already involved in some other important philosophical concepts, such as the concept of *knowledge*. Knowledge requires ideal reasoning. When one's justification for a belief can be defeated by better reasoning, such a belief does not count as knowledge. The point here is that although clarifying the idea of ideal reasoning might be difficult, the idea itself is not a foreign one. For the sake of our discussion here, it is better to simply adopt the intuitive notion of ideal reasoning. Still, it seems like we can never be certain whether something is ideally conceivable, as it seems unlikely that humans are even capable of using ideal reasoning. The perfect rational being might not exist. However, it is arguable that we can get close sometimes, just like we can be pretty certain that "1+1=2" is not going to be defeated by any better reasoning. Furthermore, whether we can ever achieve ideal reasoning seems to be a different question to whether the ideas of ideal reasoning and ideal conceivability make sense. With the intuitive notion of ideal reasoning, we seem to be able to generate a decent understanding of the idea of ideal conceivability at the very least.

The third dimension is rooted in the two-dimensional semantics framework. The primary/secondary distinction in conceivability is better explained using Chalmers's (2002) own words:

> We can say that S is *primarily conceivable (or epistemically conceivable)* when it is conceivable that S is actually the case. We can say that S is *secondarily conceivable (or subjunctively conceivable)* when S conceivably might have been the case [given how things actually are]. (p. 157)

The two conceivability correspond to two ways of thinking about a world or a scenario. When we conceive of a world as actual, primary conceivability is in play, and when we conceive of a world as merely counterfactual, secondary conceivability is in play. For example, given that water is H_2O in the actual world, "water is XYZ" is primarily conceivable but secondarily inconceivable. We can certainly conceive of a world where the local watery stuff is XYZ. However, this will be a world where water is XYZ only if this world is considered as actual. When this world is considered as counterfactual, "water is XYZ" remains false in this world.

Primary conceivability is always *a priori*, while secondary conceivability is often *a posteriori*. This is because when we deal with secondary conceivability, we always need to take into account the empirical facts of our world. In this thesis, I will mostly focus on primary conceivability.

1.4.3.2 The Conceivability-Possibility (CP) Thesis. For the zombie argument to work, we want the conceivability of zombies to entail, or at least act as a good guide to, the possibility of zombies. This requires us to bridge the gap between conceivability and possibility. So, we want to demonstrate the truth of the conceivability-possibility (CP) thesis: If something is conceivable, it is also possible.

With different kinds of conceivability and possibility, we can generate different CP theses. First of all, it is clear that all CP theses concerning *prima facie* conceivability can be rejected, as *prima facie* conceivability can potentially be defeated by better reasoning. Even scenarios containing incoherences or contradictions might turn out to be *prima facie* conceivable. Second, primary conceivability does not seem to support secondary possibility, and secondary conceivability does not seem to support primary possibility. If the possible worlds involved are considered as actual, then primary conceivability and possibility are of concern; if the possible worlds involved are considered as counterfactual, then secondary conceivability and possibility are of concern. In our case, we want to focus on primary conceivability and primary possibility, and there are two CP theses that are most likely to be true:

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

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

According to Chalmers, CP+ is almost certainly true, and CP- is very likely to be true. Since positive conceivability entails negative conceivability, CP- entails CP+. One can potentially bridge the gap between CP+ and CP- by arguing that negative conceivability entails positive conceivability, which Chalmers calls “the NEGPOS thesis.” The truth of NEGPOS requires the denial of general inscrutabilities and open inconceivabilities. I will not get into the details of NEGPOS here as they are not the most relevant to our discussion.

The important thing is to demonstrate why either CP+ or CP- is true. Chalmers focused on CP+ in his paper and pointed out that counterexamples to CP+ must involve cases of strong metaphysical necessities (or strong necessities for short). Strong necessities exist when we can ideally conceive of a situation (as actual) where a statement is falsified, despite the statement being primarily necessarily true. Chalmers has provided rather detailed defences against putative examples of strong necessities in both his (2002) paper *Does Conceivability Entails*

Possibility? and (2009) paper *The Two-Dimensional Argument against Materialism*. I will not repeat the details here. More about strong necessities will be discussed later in Chapter 2.

Despite CP+ being the more defensible thesis, I think CP- is also very plausible. One advantage of focusing on CP- instead of CP+ comes to negative conceivability being a more straightforward concept than positive conceivability. As both the concepts of negative conceivability and logical possibility are based on the idea of coherence, it is not so difficult to draw the link between conceivability and possibility. Moreover, by relying on the idea of coherence, we can avoid defining conceivability in terms of possibility, which most proponents of the CP thesis want to avoid. Again, I will defend CP- more thoroughly in Chapter 2.

1.5 Consciousness

The main goal of the zombie argument is to argue against physicalism (especially structural physicalism – the kind of physicalism that claims structural and dispositional information tells us everything about our world) by arguing that consciousness fails to metaphysically supervene on the physical. The way we understand the concept of *consciousness* is critical to how we evaluate the argument. In this section, I will introduce the common conceptions of consciousness, and pick out the one that matters to our argument.

1.5.1 What is Consciousness?

Consciousness is one of the most discussed topics in philosophy of mind, and maybe in all of philosophy. Ever since the development of contemporary science and psychology, we have gained plenty of knowledge about our mental affairs. Psychologists usually first try to connect mental states with certain behaviours, before attempting to discover the corresponding physiological states (usually brain states) that cause these behaviours. Plenty of success has been achieved by contemporary psychology and neuroscience. Still, consciousness remains one of the most difficult puzzles in the study of the mind. Later in this section, I will also explain why I think the modern scientific methods have trouble truly demystifying consciousness.

Let us go back to the main question: What is consciousness? There are a variety of ideas associated with this term, such as wakefulness, reportability, self-consciousness, voluntary control, access consciousness, and phenomenal consciousness. These six conceptions of consciousness do not exhaust all the ideas people associate with the term "consciousness," but they are the most commonly used ones (for other introductions to the varieties of consciousness, see Lycan, 1996; Chalmers, 1996; Van Gulick, 2014).

The concept that we are interested in here is phenomenal consciousness. Simply put, phenomenal consciousness concerns experience. There is something that makes phenomenal consciousness stand out among other kinds: All other kinds of consciousness mentioned above are accessible and measurable from a third-person perspective, contrary to the private and subjective nature of phenomenal consciousness. Still, can we do a better job at clarifying what the concept of *phenomenal consciousness* is? Defining "consciousness" in terms of other concepts is an extremely difficult task, if not impossible (Chalmers, 1996; Ludwig, 2003). This is most likely because the concept of *phenomenal consciousness* is primitive and irreducible. The closest definition we can get is probably "the subjective character of experience," but then, we will need to define the term "experience," which is equally problematic.

Instead of defining it in terms of other concepts, philosophers often clarify the idea by describing the phenomenon. In Nagel's (1974) words, for an organism to be phenomenally conscious, there must be "something it is like to be that organism" (p. 436). In Jackson's (1982) words, phenomenal consciousness is like "the hurtfulness of pains, the itchiness of itches, pangs of jealousy… the characteristic experience of tasting a lemon, smelling a rose, hearing a loud noise or seeing the sky" (p. 127). In Chalmers's (1996) words, "consciousness experiences range from vivid colour sensations to experiences of the faintest background aromas; from hard-edged pains to the experience of thoughts on the tip of one's tongue…" (p. 4).

Although it is difficult to pin down the concept of *phenomenal consciousness*, conscious experiences should easily be our most familiar thing. It is one of the few things, if not the only thing, that we have direct access to. It is safe to say that phenomenal consciousness is real, ineffable, private, and subjective. Denying any of the above features carries the risk of misinterpreting what phenomenal consciousness really is. Of course, certain physicalists might argue that such a concept does not exist or is incoherent. I will discuss more in Chapter 2.

"Consciousness" is not the only term associated with our concept. Other terms include "experience," "qualia," "raw feels," "phenomenal," and so on. These terms do not always refer to the same concepts, but sometimes they do. In this thesis, I consider them all to refer to phenomenal consciousness by default. Furthermore, I will be using the term "consciousness" for phenomenal consciousness by default. Clarifications will be provided in places where the differences between these terms do matter.

1.5.2 Two Concepts of Mind

To properly understand our concept of *phenomenal consciousness*, there is a distinction that deserves special attention: the distinction between the phenomenal concept of mind and

the psychological concept of mind. This distinction was most notably proposed by David Chalmers (1996), but he is not the only one who finds this distinction genuine and intuitive. This distinction is crucial to our discussion. Objections to the zombie argument can be easily constructed when the psychological concept is used, but these objections can also be easily defused since they are simply cases where the wrong concept is involved.

To understand this distinction, suppose you accidentally touch a hot stove. You quickly withdraw your hand, display expressions of aversion, and scream "ouch." At the same time, according to contemporary neuroscience, a bunch of neurons in your body are activated, and your brain structure is slightly modified by these neurological inputs. However, besides all these behavioural and neuro-biological events, you have some specific experiences – it hurts. In this case, the psychological concept picks out the behavioural and neuro-biological events, and the phenomenal concept picks out the pain experience.

The distinction becomes more obvious when we observe everything from a third-person perspective. Suppose it is not you but your friend Bob who touches the stove. You observe all the behaviours, and if you have all the appropriate machinery, you could observe the firing of nerve fibres and activations of certain brain areas. However, there is no way for you to feel Bob's conscious experiences (if he has any). Here, most people will agree that Bob is in pain. But is Bob in psychological pain or phenomenal pain? Probably both, but we can only be certain of the existence of psychological pain. Since the psychological concept of *pain* is based on behaviours and neural-biological states, the behaviours and neural-biological states directly entail the existence of psychological pain. On the other hand, with phenomenal pain, we can only infer its presence using evidence from behaviours and neural-biological states. We do not have direct access to Bob's phenomenal pain.

If this distinction between the two concepts is genuine, we can call the properties picked out by the psychological concept "psychological properties" and the properties picked out by the phenomenal concept "phenomenal properties." It is the phenomenal properties that we care about here in our thesis. However, acknowledging the distinction between the two concepts does not mean giving up physicalism. Many physicalists publicly acknowledge that there are two distinct concepts of mind, but they think that the two sets of concepts share the same referents (see Smart, 1959; Loar, 1990; Block & Stalnaker, 1999; Papineau, 2002). Of course, there are also philosophers who refuse to consider this distinction to be genuine, mostly by arguing that the only concepts of consciousness we can have are the ones that are causally or functionally defined (Lewis, 1966, 1995; Dennett, 1993). More on this will be discussed in Chapter 2.

In everyday life, the distinction here is not significant. It seems like laypeople often mix up the two concepts. This is probably due to the regular co-occurrences of the phenomenal properties and psychological properties that can be observed from the first-person perspective. For example, my behaviour of saying "ouch" is very often accompanied by my phenomenal pain, which makes me associate them with each other. We develop plenty of these associations and tend to generalise them to other people. This at least makes *prima facie* sense considering the physical similarities between humans. Furthermore, contemporary psychology and neuroscience aim to develop correlations between specific neural-biological activities and experience-related behaviours. With these correlations, scientists are then able to propose links between neural-biological states and phenomenal consciousness. This then leads to the principle of coherence: Phenomenal events are always accompanied by specific psychological events (Chalmers, 1996).

The principle of coherence is quite likely to be true in the actual world, considering the current empirical findings we have. However, I do not think it enables psychology and neuroscience to completely demystify consciousness. To see why, let us re-examine how the link between psychological states and phenomenal states is established. Scientific research on consciousness first relies on first-person associations, such as connecting the event of being hit with the experience of pain and connecting the expression of smiling with the experience of happiness. Since we only have direct access to our own experiences, these associations can only be generated from the first-person perspective. Then, these associations get generalised to other people. Whenever someone else is getting hit, we infer that they are experiencing pain, and whenever someone is smiling, we infer that they are experiencing happiness. Neuroscience then connects these behaviours (such as smiling, getting hit, and verbal reports) with specific neural-biological states and claims those states as the neural correlates of consciousness. This research method has become very popular nowadays, and findings in neural correlates of consciousness no doubt give us plausible knowledge about the correlations between psychological states, which are physically defined, and phenomenal states (for relevant studies, see Engel & Singer, 2001; Rees et al., 2002; Lamme, 2006; Koch et al., 2016).

The problem is that the link between psychological states and phenomenal states presented here is not as rigid as what some physicalists might want it to be. There are two potential gaps left to be closed. First, we develop the associations between physical events (including behaviours) of myself and phenomenal experiences via the observations of limited co-occurrences of them. After all, these are mere correlations. It neither automatically makes the two things identical to each other nor entails some deep connections between the two.

Second, and more importantly, the correlations between physical events and phenomenal experiences of others are based on mere inferences. We do not have direct access to phenomenal experiences that we do not have, but we nonetheless infer their existence. For example, most of us believe dogs and cats and fellow humans have some sort of phenomenal experiences, even though we do not feel their experiences. We think so because of their similarities in behaviours to us, but the assumption that behaviours of such-and-such suggesting experiences of such-and-such is too big to be taken for granted. What these two gaps show is that although contemporary science has been doing a great job explaining psychological consciousness, it says very little about the link between psychological consciousness and phenomenal consciousness. We cannot simply assume that the two share some deep metaphysical connections.

1.6 Physicalism and Russellian Monism

Physicalism is the target of the zombie argument. So, both defending and objecting to the argument will depend on what physicalism is. Making physicalism either true or false would be easy if we could just alter the meaning of the term "physicalism" freely, but that is clearly not what philosophers want. For there to be a non-superficial debate, we need to clarify the idea of physicalism in play. There are two core questions here: What the formulation of physicalism is, and what it means to be physical. In this section, I will first provide the relevant formulation of physicalism. Then, I will focus more on the concept of *physical*. Last but not least, I will discuss the view of Russellian monism, a view that is seen as a kind of physicalism sometimes but nonetheless differs significantly from traditional physicalism.

1.6.1 The Formulation of Physicalism

Physicalism, in its simplest form, is the ontological thesis that everything is physical. Sometimes physicalism is also known as "materialism." Physicalism is often considered as the successor of materialism, and therefore differs from materialism in some subtle ways (Ney, 2008; Stoljar, 2022). Many earlier literatures only use the term "materialism," even though their ideas are extremely similar or identical to the contemporary physicalism (Armstrong, 1968; Levine, 1983; Lewis, 1995). For the sake of simplicity, I will use these two terms interchangeably in this thesis.

Let us go back to the phrase "everything is physical." I will deal with the term "physical" in more detail later, but even if we adopt the intuitive usage of the term, it does not immediately seem like everything is physical. It might be uncontroversial to consider things

such as rocks, water, and atoms as physical, but it becomes more debatable when we consider things such as living organisms, thoughts, complexity, knowledge, and moral values. At the very least, things in the latter group do not appear as intuitively physical, and physicalists probably do not want to simply assume that they are physical without reason. One way to get around it is to say that the existence and distribution of these things, such as knowledge and moral values, are fixed by or dependent on the existence and distribution of things that appear obviously physical, such as rocks and atoms. One example of formulating physicalism in this way comes from Jackson's (1998) *entry by entailment thesis*. What Jackson notices is that what physicalism needs is for any story told in apparently non-physical terms to be entailed by the complete physicalist account.

The entry by entailment thesis relies on the idea of supervenience. So, even though there are things that do not seem to be physical at first, as long as the information about them supervene on the physical information, physicalism is fine with their existence. For example, learning about the distribution of every single particle (or whatever entities are used in the fundamental physical theory) within an organism automatically tells us about its complexity, possession of knowledge, emotions, biological fitness, etc.[16] If such supervenience can be achieved, the complete set of physical information about our world fixes all other information about our world. With this idea in mind, Jackson (1998) proposes the following formulation of physicalism:

> Physicalism is true in our world if and only if any world which is a minimal physical duplicate of our world is a duplicate *simpliciter* of our world.

This formulation is also similar to the one given by David Lewis (1994). The core idea here is the supervenience of all things on the physical. Once we duplicate all the physical features of our world, we get a duplicate *simpliciter* of our world. There are other ways of formulating physicalism, such as relying on the ideas of realization, grounding, and fundamentality, but it should be uncontroversial that all of them require some kind of supervenience (for discussions, see Lewis, 1994; Ludwig, 2003; Melnyk, 2003; Levine, 2020; Stoljar, 2022). In other words, supervenience might not be sufficient for physicalism, but it is most likely necessary. This also means that to defeat physicalism, we simply need to defeat supervenience.

Sometimes physicalists disagree on whether physicalism requires all positive facts about our world to be entailed by the complete physical story of our world. If all properties

[16] Technically, we still need to know about the bridge laws between physical properties and apparently non-physical properties. Physicalists have to argue that these laws are considered as physical as well.

logically supervene on physical properties, then it seems like knowing the complete physical story will automatically tell us everything about our world. In this case, the totality of all physical information about our world *a priori* entails all information about our world. This is the view *a priori* physicalists adopt (for discussions and support, see Lewis, 1994; Chalmers, 1996; Jackson, 1998; Chalmers & Jackson, 2001; Kirk, 2005). On the other hand, some physicalists do not consider all properties to logically supervene on physical properties. Instead, they think that all the apparently non-physical properties are connected to the physical properties in a deep metaphysical way. They prefer saying that the physical information merely necessitates, but does not entail, the other information about our world. In this case, knowing the totality of physical information does not tell us the totality of all information. This is the view *a posteriori* physicalists adopt (for discussions and support, see Block & Stalnaker, 1999; Hill & McLaughlin, 1999; Loar, 1999; Papineau, 2002; Levine, 2020).

A priori physicalism is arguably the stronger view here, which means it is more restricted and demands more support. If the totality of physical information *a priori* entails all information, then all properties logically supervene on physical properties. Logical supervenience guarantees physicalism. The motivation for *a posteriori* physicalism appears to be avoiding the dependence on logical supervenience. *A posteriori* physicalism allows the failure of logical supervenience, as long as there is still metaphysical supervenience. Chalmers (2002, 2003a) distinguishes between type-A and type-B materialism, where type-A materialism resembles *a priori* physicalism, and type-B materialism resembles *a posteriori* physicalism. The two kinds of physicalism (materialism) pose quite different objections to the zombie argument, which means defending the zombie argument against them requires different strategies.

1.6.2 The Concept of Physical

Even if we can come up with a definite formulation of physicalism, there remains the question of what it means for something to be physical. This is not an easy question to answer. Philosophers are far from reaching a consensus on the concept of *physical*, yet such a concept is essential to our understanding of physicalism. There are three major conceptions of physical (Stoljar, 2022): the theory-based conception, the object-based conception, and the Via Negativa conception.

The theory-based conception is arguably the most popular one out of the three, and it characterizes the physical using the physical theory. According to this conception, a property

is physical if and only if the physical theory tells us about such a property.[17] For example, since the current physical theory tells us about the property of having mass and the property of having electric charges, the properties of having mass and having electric charges count as physical properties. Properties that are picked out by the theory-based conception can be called "t-physical properties."

The most discussed problem with this conception is what most philosophers refer to as Hempel's dilemma (Hempel, 1969, 1980; Stoljar, 2010; Tiehen, 2018). The key idea is the following: If physicalism is defined using the physical theory, what specific physical theories should we use? If we use the current physical theory, then physicalism is most likely to be false since the current physical theory is probably incomplete. But if we use a future or ideal theory, then the truth of physicalism is trivial since we have no idea what kind of theory it might be. Such a future theory might even contain mental terms and properties. The question then becomes why we should still call such a theory physical. Crane and Mellor (1990) even argue that because of Hempel's dilemma, there is no meaningful formulation of physicalism. Currently, the most promising response seems to be adopting a modified version of futurism, which restricts what a future physical theory can include (see Ludwig, 2003; Dowell, 2006; Ney, 2008).

Instead of relying on our intuitions about the physical theory, the object-based conception relies on our intuitions about physical objects.[18] The object-based concepts pick out the properties that are possessed by our paradigmatic physical objects such as rocks, tables and water. For example, if some or all of our paradigmatic physical objects are solid and extended in space-time, then the properties of being solid and extended in space-time count as physical under this conception. Of course, the properties of being solid and extended in space-time also count as t-physical properties, which means the o-physical conception picks out at least some, if not all, t-physical properties. However, the o-physical concepts also pick out certain non-t-physical properties.

To understand what extra properties the o-physical conception picks out, we need to first understand the categorical/dispositional distinction. It is commonly thought that the

[17] One might worry that the conception here is circular, as we can easily conceptualise physical theories as the theories that tell us about the physical properties. To avoid this issue, we can simply rely on our intuitive judgements on physical theories without using any specific definition. We seem to intuitively judge certain theories as physical, and reach a pretty good consensus on this subject matter.

[18] The object-based physical conception here is similar to the broad notion of physical Chalmers (2015a) raises.

physical theory only picks out dispositional properties, not categorical properties (see Chalmers, 1996, 2009; Chalmers & Jackson, 2001; Alter & Nagasawa, 2012). Take the property of having mass as an example. The concept of *mass* is based on the concepts of resistance and weight, in a way that we only know about the mass of something by feeling its weight or resistance to force. All other t-physical properties are similar, as their concepts are all based on our interactions with the carriers of the properties, which makes the t-physical properties dispositional. There is another popular view, which is discussed by philosophers such as Bertrand Russell (1927) and Simon Blackburn (1990), that considers dispositional properties must have categorical grounds and that the fundamental properties of our world are categorical. What follows is that physical theory either only tells us about the dispositional non-fundamental properties or about what the fundamental properties do without telling us what they really are. Either way, we have very limited access to the intrinsic nature of our world (for discussion, see Jackson, 1998; Lewis, 2001).

The extra properties that the o-physical concepts pick out are the intrinsic categorical properties of the paradigmatic physical objects. However, there is a catch: even though we can pick out these categorical properties with the o-physical conception, we still do not really know what they are. What the o-physical concepts do is simply picking out whatever categorical intrinsic properties that ground certain dispositional properties or play certain dispositional roles. This naturally leads to the debate of whether these o-physical properties can truly be called "physical," which I will talk about later.

The last one is the via negativa conception. The most famous advocate of this view is probably David Papineau (Spurrett & Papineau, 1999; Papineau, 2002; Montero & Papineau, 2005). To use Papineau's (2002) own words, the term "physical" means "non-mentally identifiable and inanimate." Philosophers usually interpret this view as the following: A property is physical if and only if it is not mental. One problem with the via negativa conception is the possibility of properties that are non-mental and non-physical at the same time, with one example being vitality. Another obvious problem is that it contradicts the identity theory that most physicalists want to adopt, as many physicalists want to say that phenomenal concepts and physical concepts co-refer (Stoljar, 2010; Tiehen, 2018). I will not focus too much attention on this view here, as the other two conceptions are much more essential to the two-dimensional zombie argument.

1.6.3 Russellian Monism

With the t-physical and o-physical conceptions, we can create t-physicalism and o-physicalism. The former is what most philosophers think of when the term "physicalism" is mentioned. However, o-physicalism is by no means a completely novel view. It closely resembles a view that has existed for almost a century: Russellian monism. Most of the discussions on Russellian monism nowadays can be traced back to Bertrand Russell's (1927) *The Analysis of Matter*, with some of the more contemporary proponents of the view being Grover Maxwell (1978), Michael Lockwood (1989), and David Chalmers (2015a).[19] There seems to be a resurgence of interest in this view over the last few decades.

Put simply, Russellian monism can be understood as the ontological view where only one fundamental kind of properties, which are called the "quiddities" or "inscrutables," exists in our world. These quiddities are the referents of both our phenomenal and o-physical concepts, even though their natures are not directly captured by any physical concepts. According to Alter and Nagasawa (2012), Russellian monism has three core theses: structuralism about physics, realism about quiddities, and (proto)phenomenal foundationalism. Structuralism about physics, as I discussed earlier, claims that physics only tells us about dispositional or extrinsic properties. Realism about quiddities claims that there are quiddities in our world that are not dispositional or extrinsic. (Proto)phenomenal foundationalism claims that phenomenal or protophenomenal properties constitute at least some of these quiddities. In the following, I will explain more details about these theses.

There are three notes concerning Russellian monism I want to talk about here. First, one of the core ideas of Russellian is the distinction between properties revealed by physics and the fundamental properties that are not revealed by physics. Proponents of Russellian monism usually describe the former properties as dispositional, extrinsic, relational, and structural-and-dynamic and describe the latter as categorical, intrinsic, non-relational, and non-structural-and-non-dynamic (Alter & Nagasawa, 2012). Four contrasts (e.g., dispositional vs. categorical) can be created here. I will not get into the differences between these contrasts here as the differences themselves are very subtle. However, the distinctions involved in each individual contrast are rather undebated. Terms in the first set (dispositional, extrinsic, relational and structural-and-dynamic) can often be used interchangeably, so can the terms in the second set (categorical, intrinsic, non-relational, and non-structural-and-non-dynamic). To simplify the matter, I will consider terms within each set to share the same meaning by default.

[19] Chalmers (2003a) sometimes also calls it "type-F monism."

Physics only reveals the dispositional aspect of our world, as it purely relies on our interactions with the world. As a result, physics only tells us about what the world has done to us, without revealing its intrinsic nature.

The second note concerns the different types of Russellian monism. Russellian monism considers the only kind of fundamental entities that exists in our world to be quiddities. But what are quiddities? With different views on the nature of quiddities, different kinds of Russellian monism can be created. If quiddities are seen as phenomenal properties, Russellian monism resembles a kind of idealism (see Russell, 1927); if quiddities are seen as neutral properties such as protophenomenal properties, Russellian monism resembles a kind of neutral monism (see Chalmers, 2003a, 2015a); if quiddities are seen as a special kind of physical properties, Russellian monism resembles a kind of physicalism (see Stoljar, 2001a; Papineau, 2002). David Chalmers himself endorses panprotopsychism, which makes his Russellian monism a kind of neutral monism. Although I agree with most of Chalmers's views, this will be the one that I am uncertain about, mainly due to the vagueness of how protophenomenal properties and phenomenal properties connect. However, I will not attempt to provide an objection to Chalmers's panprotopsychism here as it does not directly serve the purpose of the thesis.

The last note concerns panpsychism and panprotopsychism. Since quiddities are considered as the sole kind of properties that constitutes the fundamental nature of our reality, they are ubiquitous. If we consider quiddities to be phenomenal properties or protophenomenal properties, what follows is that phenomenal properties or protophenomenal properties are everywhere in our world. Many philosophers find panpsychism counterintuitive or even repugnant, but there have been some solid defences for panpsychism and protopanpsychism recently (Chalmers, 2015a; Goff, 2017). Once again, due to the word limit of the thesis, I will not attempt to provide any further support for panpsychism here, but I do consider panpsychism a coherent and feasible view.

It might be useful to keep in mind the three notes that I mentioned above, as they are complications associated with Russellian monism. If we do wish to endorse the two-dimensional zombie argument, Russellian monism will be one of the two possible outcomes, which makes the complications above relevant. However, I think the most important idea about Russellian monism here is how it refers to quiddities. Russellian monism does consider physical concepts and phenomenal concepts to refer to the same set of properties. However, the physical concepts merely reveal the extrinsic but not intrinsic aspects of the properties. This is similar to the case where we pick out H_2O with the concept of *water*. The concept of *water*

does refer to H_2O in our world, but only via its extrinsic aspects, and because of this, the primary intension and secondary intension of the concept differ. Two worlds can share the same extrinsic appearance but share different intrinsic natures at the same time. In cases like this, what the concept of *water* picks out in these worlds will then depend on whether these worlds are considered as actual or counterfactual. Following this idea, for physical concepts and phenomenal concepts to co-refer, the physical concepts involved must also have distinct primary and secondary intensions. O-physical concepts do exactly that. Two possible worlds might be extrinsically identical but intrinsically different. Physics will tell us that these two worlds are identical, but proponents of Russellian monism will argue that they are not, even though there is no observable way of telling them apart. This is also why Russellian monism is not often considered to be a kind of physicalism, as it drastically differs from traditional physicalism.

1.7 The Two-Dimensional Argument Clarified

After going through the key concepts involved in the two-dimensional zombie argument, it is time to examine the argument in greater detail. In this section, I will go through the premises and the conclusion one by one and explain why the argument is formulated in the way it is. After examining the argument, it should become clear what the real target of the argument is and how the argument attempts to bring it down.

1.7.1 Premise 1: Zombies are Ideally Primarily Conceivable

As mentioned before, zombies are physical duplicates of humans that lack phenomenal consciousness, but what concept of *physical* is involved here? The argument here can be run in two ways, depending on whether o-physical concepts or t-physical concepts are chosen. We get o-zombies if we use o-physical concepts, and t-zombies if we use t-physical concepts. O-zombies share the same o-physical properties with us, while t-zombies merely share the same t-physical properties with us. If we focus on this premise alone, it does not matter too much which concepts we choose, since proponents of the argument will consider both types of zombies ideally primarily conceivable. The difference appears when we consider the secondary conceivability and possibility of these two types of zombies. If we take the later premises into account, then using the o-physical concept makes more sense.

Although zombies do not have phenomenal consciousness, they are identical to humans in terms of psychological consciousness. This is because psychological consciousness is defined in terms of behaviours and neural-biological activities, and both are commonly

considered to logically supervene on the physical. So, zombies behave like their human twins, and they also share identical biological structures. What zombies lack are merely the subjective experiences, or qualia, or raw feels.

It is important to note that both phenomenal concepts and t-physical concepts have identical primary and secondary intensions,[20] while o-physical concepts have distinct primary and secondary intensions. Interestingly, Chalmers (2009) himself thinks that the argument works even if phenomenal concepts have distinct primary and secondary intensions; we simply need to formalise premises 3 and 4 differently. However, I think it is fairly safe to say that phenomenal concepts are transparent and pick out their referents directly, which makes their primary and secondary intensions identical. To simplify the matter here, I will take this as an assumption.

Some versions of the argument use zombie worlds instead of zombies, but there are complications with the use of zombie worlds. A full-blown zombie world is a physical duplicate of our world with no conscious agent at all. There is no way of experiencing this world from within, which might contribute to the difficulty of conceiving of such a world. More specifically, some people might dislike the idea of conceiving a world in which their counterparts are not conscious. I do not want to get into the complication in too much detail, but just to be safe, we can choose to conceive of a world with merely one or a few zombies (that are not ourselves). These worlds, which I shall call "partial zombie worlds," are enough to break the supervenience of consciousness on the physical, as almost all physicalists deny the existence of any zombie in our world. The possibility of a world that is physically identical to ours but phenomenally different, even in a small way, will bring down physicalism.

In some of the more recent literature, the idea of zombies and zombie worlds is sometimes written as P&¬Q, where P is the conjunction of all (micro)physical facts about our world and Q is an arbitrary phenomenal fact (Chalmers, 2009). P&¬Q can either represent a complete zombie world, or any partial zombie world, or even an inverted world (where some conscious agents have different conscious experience from their counterparts in our world), depending on what arbitrary phenomenal fact Q represents. For the sake of consistency, I will keep using the term "zombie" in this thesis.

[20] To clarify, I am not saying that phenomenal concepts and t-physical concepts share the same primary (or secondary) intensions. What I am saying is that the primary and secondary intensions of the phenomenal concepts are identical, and the primary and secondary intensions of the t-physical concepts are also identical.

Let us now turn our attention to the conceivability part. Ideal conceivability is used since *prima facie* conceivability clearly acts as an inferior guide to possibility. Primary conceivability is used since secondary conceivability involves the use of *a posteriori* empirical information, which will be dealt with in premises 3 and 4. Moreover, primary conceivability is easier to achieve compared to secondary conceivability due to fewer restraints. The argument here intends to first establish the easier-to-achieve primary conceivability and primary possibility, before moving on to secondary possibility in later premises.

Chalmers (2002, 2009) himself does not specify whether positive or negative conceivability should be used in premise 1. It seems to me that Chalmers thinks the argument works with both kinds of conceivability. We will later see a subtle balance between premise 1 and premise 2: there seems to be stronger support for premise 1 if we use negative conceivability, but stronger support for premise 2 if we use positive conceivability. Personally, I prefer using negative conceivability since it is an easier concept to explain, and the definition of the concept has a clearer link to the concept of *possibility*. Without explicit clarifications, I will be using negative conceivability by default in this thesis.

1.7.2 Premise 2: If Zombies are Ideally Primarily Conceivable, Zombies are Primarily Possible

This premise is the conceivability-possibility thesis (CP thesis), which intends to bridge the gap between conceivability and possibility. If the CP thesis is true, this premise is true. As the primary conceivability of zombies is involved here, the premise only cares about the primary possibility as the consequent. But what does it mean for zombies to be primarily possible? For zombies to be possible is for zombies to exist in at least some possible worlds, and for zombies to be primarily possible is for zombies to exist in at least some possible worlds when the possible worlds are considered as actual. In our case, it means there must be at least one possible world that is dispositionally and structurally identical to our world but phenomenally different.

With the option of using negative or positive conceivability, there are two ways the premise can be formulated, with one being CP+ and the other being CP-. Chalmers seems to be endorsing both CP+ and CP-. As I mentioned before, there is a subtle balance between premise 1 and premise 2 in terms of whether positive or negative conceivability is involved. Since CP+ is considered as the weaker thesis, the use of positive conceivability will make premise 2 easier to be true. At the same time, the use of negative conceivability will make

premise 1 easier to be true. And as I said earlier, negative conceivability will be used by default in this thesis.

1.7.3 Premise 3: If Zombies are Primarily Possible, Zombies are either Secondarily Possible or Russellian Monism is True.

Premise 3 specifies the two options one can adopt once they accept that zombies are primarily possible: to accept that zombies are also secondarily possible or to accept Russellian monism. Here, secondary possibility resembles the metaphysical possibility most philosophers talk about. To further clarify the premise, we can deconstruct it into two separate claims:

(P3a) If zombies are primarily possible, then zombies are either secondarily possible or secondarily impossible.

(P3b) If zombies are secondarily impossible, Russellian monism is true.

(P3a) is obviously true, as the consequent contains two events that are mutually exclusive and collectively exhaustive. Either zombies are secondarily possible or secondarily impossible, with no third option available. Since the primary and secondary intensions of t-physical concepts are identical, if t-zombies are primarily possible, they must also be secondarily possible. O-zombies, on the other hand, can be primarily possible without being secondarily possible since the primary and secondary intensions of o-physical concepts are different.

Why is that the case? Consider the claim "water is XYZ" again. This claim is primarily possible but secondarily impossible. This is because there are possible worlds where the local watery stuff is XYZ. When these worlds are considered as actual, the concept of *water* picks out the XYZs, and when these worlds are considered as counterfactual, the concept of *water* does not pick out the XYZs. It is a similar story when it comes to zombies. If o-physical concepts and phenomenal concepts co-refer in our world, and there are possible worlds that are dispositionally identical but phenomenally different to our world, zombies will be primarily possible without being secondarily possible. Here, the o-physical concepts in our world either pick out phenomenal properties or whatever properties that ground the phenomenal properties. The primary intension of o-physical concepts might pick out all kinds of other properties in other possible worlds, but the secondary intension always picks out phenomenal properties or the properties that ground phenomenal properties. Any world that is phenomenally different to our world will be a world physically different to our world, according to the secondary intension of our physical concepts.

On the other hand, if physical concepts and phenomenal concepts do not co-refer in our world, zombies will be secondarily possible, since the referents of physical concepts and the

referents of phenomenal concepts can easily come apart. Here, the secondary intension of physical concepts never picks out phenomenal properties or any properties that ground phenomenal properties in any possible world.

What about (P3b)? If zombies are secondarily impossible, it means that o-physical concepts and phenomenal concepts co-refer in our world. There might be a world (let us call it world X) that is dispositionally identical but phenomenally different to our world, but whatever intrinsic properties ground those dispositional properties in X will be different to the intrinsic properties that ground the same dispositional properties in our world. In other words, despite X being dispositionally identical to our world, it shares a different intrinsic nature, and if our physical concepts pick out the intrinsic nature of our world, then X is not physically identical to our world. To be able to achieve that, we need physical concepts that have distinct primary and secondary intensions, which only o-physical concepts do. Physicalism formulated using o-physical concepts closely resembles Russellian monism. So, it follows that the only way to make zombies secondarily impossible is to adopt Russellian monism.

1.7.4 Premise 4: If Zombies are Secondarily Possible, Physicalism is False

Physicalists often consider metaphysical supervenience to be a requirement for physicalism. If consciousness fails to metaphysically supervene on the physical, physicalism is false. Two-dimensional semantics simply interprets metaphysical modality as the secondary modality. So, to say that zombies are secondarily possible is to say that it is metaphysically possible to generate a world that shares identical physical structure but a different phenomenal structure. This suggests the failure of metaphysical supervenience.

We can tell a similar story with Jackson's (1998) formulation of physicalism that was mentioned before. Zombies being secondarily possible suggests that there is a possible world considered as counterfactual that is physically identical to our world (even on an intrinsic level) but still lacks some phenomenal states that our world has. It suggests that simply duplicating the physical information, even the o-physical information that includes some intrinsic nature of our world, does not automatically bring along all the phenomenal information.

The secondary possibility of zombies suggests that physical concepts and phenomenal concepts fail to co-refer in our world. This, combined with the assumption that phenomenal concepts do refer to something real,[21] suggests the failure of physicalism, as physical

[21] If there is no phenomenal consciousness in the first place, zombies are obviously secondarily possible, as they are simply identical to us. However, it also means physicalism is true in our

information fails to entail all information about our world. It also suggests that physical properties and phenomenal properties can be separated in a stronger sense, contrary to the case of water and H_2O, where the two properties can only be separated in a weaker sense.

1.7.5 Conclusion: Physicalism is False or Russellian Monism is True

If all the premises are true, this conclusion follows logically. In terms of the possibility of zombies, there are two potential outcomes: (1) Zombies are primarily possible and secondarily possible; (2) Zombies are primarily possible but secondarily impossible. The primary possibility of zombies will at least allow us to separate dispositional properties from phenomenal properties. The secondary possibility of zombies, if true, will further suggest that o-physical concepts and phenomenal concepts fail to co-refer. The secondary impossibility of zombies, on the other hand, will suggest that physical concepts and phenomenal concepts do co-refer, but this can only be achieved if o-physical concepts are involved. Physicalism is false if zombies are secondarily possible, and Russellian monism is true if zombies are secondarily impossible.

The conclusion is a disjunction due to our inability to figure out whether o-physical concepts and phenomenal concepts co-refer. As mentioned earlier, physics only tells us about the dispositional properties of our world without telling us what ground these properties. Without knowing what intrinsic and categorical properties our o-physical concepts refer to, we have no way of knowing whether they are also the properties our phenomenal concepts pick out.

The disjunction we have seems to be leaving open the option of physicalism being true and Russellian monism being true at the same time, which happens to be the view that Stoljar (2001a) takes. Chalmers himself seems a bit hesitant to say that physicalism is definitely false, as the term "physicalism" can represent a few different views depending on how we understand the concept of *physical* and the formulation of physicalism. At the same time, the only way to adopt this option seems to be to consider Russellian monism as a kind of physicalism (Chalmers, 2009). If that is the case, the dispute here mainly exists on a verbal and not metaphysical level. Personally, I prefer not to treat Russellian monism as a kind of physicalism due to its unique characteristics. Overall, we can consider the two-dimensional zombie argument here as an objection to structural physicalism, the kind of physicalism that considers

---

world since we do not need to worry about how phenomenal consciousness supervenes on the physical.

all facts to supervene on dispositional facts. Whether or not the argument here can be seen as an objection to physicalism in the broadest sense depends on how the term “physicalism” is used.

Chapter 2: Pre-existing Discussions on the Argument

Being one of the most influential arguments in contemporary philosophy of mind, the zombie argument (or the conceivability argument) has attracted plenty of discussions over the past few decades, with many of them being objections. Chalmers (2009) and Kirk (2019) have both done a summary of these objections and provided replies to some of them. Considering this is a thesis that focuses on the zombie argument, I will do a similar summary in this chapter just to cover some of the major pre-existing work that has been done on the argument.

The two main purposes of this chapter are to show the lack of knockdown objections to the zombie argument and to lay the foundation for the discussion in the next chapter. As mentioned in the introduction, one of the main goals of this thesis is to show that the zombie argument is highly defensible. To achieve that, we will need to examine the previous attempts at refuting the argument. Being one of the most influential anti-physicalist arguments, there is certainly no lack of objection to it, and many replies have also been made. Due to the sheer amount of material written about the argument, I will not be able to cover every single objection and reply. But at the very least, I will try to cover those that have attracted the most attention. The discussion on these objections and replies will also serve as useful background knowledge for Chapter 3, where I will focus on defending the argument against a few of the more novel objections. The discussion here can be used to show how those novel objections resemble or differ from other well-known objections and how my replies are connected to pre-existing replies.

In the following, I will devote a section to each premise. Since most objections focus on either the conceivability of zombies (premise 1) or the conceivability-possibility thesis (premise 2), I will spend significantly more words discussing the first two premises of the argument than the later two. Some objections might seem hard to classify, but I will still try to put them into one of the four sections. At the beginning of each section, I will briefly provide the reasons for supporting each premise. Then, I will get into the different kinds of objections that are raised and how they are replied to if they have been replied to at all. Besides summarising the pre-existing discussions on the argument in this section, I will also provide some of my own analyses and replies at certain places.

2.1 The Conceivability of Zombies

2.1.1 Reasons for Supporting the Conceivability of Zombies

One major reason for supporting the conceivability of zombies is intuition. This is especially the case for negative conceivability. Intuitively, there seems to be nothing incoherent or contradictory about a world being physically identical to ours but phenomenally different when the relevant concepts of *consciousness* and *physical* are used. For zombies to be conceivable, there must be no *a priori* entailment from the physical to the phenomenal. This seems to be the case with phenomenal consciousness and t-physical properties (for discussions on phenomenal consciousness and t-physical properties, see Chapter 1). No amount of dispositional/extrinsic information can tell us about what it is like to be a bat, even though most of us consider bats to be phenomenally conscious (Nagel, 1974). What about o-physical properties? Well, o-physical concepts also pick out their referents via dispositional/extrinsic properties. Since dispositional facts fail to *a priori* entail the relevant phenomenal facts, o-physical zombies are also conceivable, at least primarily.[22]

I am a bit more uncertain when it comes to positive conceivability, but I still think zombies are probably positively conceivable. I seem to be able to imagine a world that is extrinsically the same as ours but phenomenally different. There might be concerns about imagining a complete zombie world where no phenomenal consciousness exists, since the act of merely imagining any physical object or event requires experience. For example, when I imagine an apple, I visualise a red round shape and try to recreate the specific smells and tastes in my mind. So, one might argue that it is only possible to imagine a complete zombie world from the outside but not from the inside. This can potentially be used to argue against the conceivability of a complete zombie world. To avoid this potential issue, we can imagine a partial zombie world instead, where only one or a few zombies exist. I seem to have no trouble

[22] If we adopt phenomenal functionalism mentioned in Stoljar 2020 and Chalmers 2012, dispositional concepts and phenomenal concepts do have *a priori* links, which means dispositional facts and phenomenal facts might *a priori* entail one another. However, this is not the entailment most physicalists want. For example, phenomenal functionalism might claim that dispositional facts about an octopus tell us about the phenomenal facts of our experiences of interacting with an octopus but tell us nothing about the phenomenal facts of experiences that the octopus has itself.

imagining a partial zombie world from the inside, since the centre of the world, which is myself, is still allowed to have conscious experiences.[23]

As Chalmers (2002) mentions, conceivability is subject-relative: what appears conceivable to me might not be conceivable to others. This is certainly the case with *prima facie* conceivability. However, if conceivability is ultimately based on coherence (especially with negative conceivability), and if coherence is not subject-relative, then we seem to have good reasons to believe that ideal conceivability is also not subject-relative. Suppose a scenario is perfectly coherent and therefore ideally negatively conceivable. It might seem *prima facie* inconceivable to some thinkers, but it is only because these non-ideal thinkers are making a mistake. If a scenario is perfectly coherent, any ideal rational thinker will consider it conceivable.

The question now becomes whether there is, or even can be, any ideal rational thinker in our world. The answer is probably no, but I do not think this is a good enough reason for us to reject the ideal conceivability of everything. We know that many things are either true or false, but we can never be certain about their truth values due to us not being the ideal thinkers. Still, we are hugely confident in many things in mathematics, logic, and some other disciplines. Even if we can never be completely certain whether a claim or a theory is true or false, we are still free to support them and believe in them, given enough evidence and analysis. If there is no obvious incoherence or contradiction discovered in the idea of zombies, which I think is the case right now, I see good reasons for supporting the conceivability of zombies.

2.1.2 Objection: Analytic Functionalism and Eliminativism

To object to the conceivability of zombies, one might attempt to bridge the gap between physical concepts and phenomenal concepts. If physical facts logically (or conceptually) entail the relevant phenomenal concepts, zombies will not be ideally conceivable, as there exist contradictions in the idea. To achieve that, one can simply define "consciousness" in terms of the physical, or base phenomenal concepts on physical (especially dispositional) concepts. One can also introduce functional concepts as the mediator, then define phenomenal concepts functionally, and identify functional states as physical states.

Analytic functionalism does so by defining consciousness in terms of its functional or causal roles: mental states are just states that play such-and-such functional roles or cause such-

[23] Even with partial zombie worlds, one might still worry about our abilities to conceive of the absence of consciousness. Chalmers himself gave a brief justification to why that can be done in Section 4 of his 2009 article.

and-such other states. It is worth mentioning that functionalism itself is not strictly a physicalist view, but it has been adopted mostly by physicalists. Functionalism can be compatible with dualism if whatever plays the functional roles is not physical. However, since sensory stimuli and physical behaviours are most often considered as the inputs and outputs involved in functionalism, and the physical world is often considered to be causally closed, whatever plays the functional roles is often considered to be physical. Analytic functionalism is no exception. For analytic functionalism, a state is conscious if and only if the state produces or has the disposition to produce such-and-such behaviours. Analytic functionalism is also compatible with dualism, but only if the physical world is not causally closed. If the physical world is causally closed, which most philosophers think it is, the state being picked out by the analytic functionalist concept of *consciousness* will be physical. In most cases, such a state will be neural-biological. If so, any minimal physical duplicates of us, being physically and functionally identical to us, will also be phenomenally identical to us, as they must share all the conscious states that we have by definition. Therefore, zombies are inconceivable.

Another similar but more extreme view is eliminativism, which simply denies the existence of phenomenal consciousness. Chalmers (2003a) points out that eliminativism does not differ from analytic functionalism in any deep sense. Although eliminativism denies the existence of common-sense mental states, it does agree that certain physical states play specific functional roles and produce our behaviours. If we decide to define these physical states as mental states, then mental states do exist. The only difference between eliminativism and analytic functionalism, assuming that our world is physically closed, seems to be whether certain physical states should be called "mental states" or "consciousness." For some people, it might still seem controversial to say that eliminativism and analytic functionalism are fundamentally the same view. However, the uncontroversial bottom line is that both views do not consider us to be conscious in any non-functionally defined sense.

The most notable philosophers who adopt the above views are probably Lewis (1966), Dennett (1993; 1996), and Braddon-Mitchell and Jackson (1999; 2006). Sydney Shoemaker (1999) also provides a defence for functionalism in reply to the zombie argument, arguing that similarities and differences in qualitative experience are functionally definable even if individual qualia are not. Chalmers (1999) replies to Shoemaker simply by denying that qualitative similarities and differences are functionally definable. With analytic functionalism and eliminativism in general, Chalmers (2003a) sometimes calls them "type-A materialism." Chalmers points out that the debate between type-A materialists and zombists ultimately comes down to intuition. The question here is whether there is a kind of phenomenal consciousness

that cannot be functionally analysed. The zombists say yes, while the type-A materialists say no. There is arguably this strong intuition shared by many people that explaining everything physical does not automatically give us the phenomenal facts, and type-A materialism denies this intuition. As a result, type-A materialism remains a controversial view, even among the physicalists.

2.1.3 Objection: Zombies are Prima Facie but not Ideally Conceivable

Some might argue that zombies are only *prima facie* conceivable but not ideally conceivable. This means that although there is an apparent epistemic gap between physical facts and phenomenal facts, it can be closed in the ideal situation. This will explain our strong intuition about the conceivability of zombies, even though zombies are inconceivable at the limit. Chalmers (2003a) calls this view "type-C materialism," and he also points out that it is a view adopted by some influential philosophers such as Nagel (1974), Churchland (1997), van Gulick (1993), and McGinn (1989).

Chalmers considers type-C materialism an unstable view, in the sense that it can easily collapse into other views. Type-C materialism, in general, is silent about why zombies are not ideally conceivable but *prima facie* conceivable. There are mainly two possibilities for why this is the case: either that we do not currently possess ideal reasoning, or that we do not currently have access to the complete set of physical truths. If the former is true, zombies are inconceivable because all phenomenal facts will be *a priori* entailed by dispositional (or t-physical) facts after ideal reflections, and type-C materialism collapses into type-A materialism. If the latter is true, zombies are inconceivable because all phenomenal facts will be *a priori* entailed by some sort of physical facts after we have achieved the complete physics, and type-C materialism collapses into type-B materialism or Russellian monism (discussions on these two views will be provided later). Although some people might consider this instability an advantage, it is hard to deny how this view will ultimately collapse into one of the other views when we try to examine the reason for our *prima facie* conceivability more closely. For this reason, objections to the other views can sometimes be used to argue against type-C materialism.

More recently, Worley (2003) and Bailey (2007) also pose their objections to the zombie argument by pointing out the gap between *prima facie* and ideal conceivability. Both Worley and Bailey question the idea of ideal conceivability, saying that humans, as non-ideal thinkers, cannot truly know whether zombies are ideally conceivable or not. In reply to Worley and Bailey, Chalmers (2009) argues that despite us being non-ideal thinkers, there are still

things that we can be very confident of, such as "0+1=2" not being ideally conceivable and "someone exists" being ideally conceivable. Given enough thinking and reasoning, we can at least get close to determining whether something is ideally conceivable. So, although we might not be absolutely certain that zombies are ideally conceivable, we have at least good reasons to believe in it if we cannot find any incoherence in the idea after scrutiny.

2.1.4 Objection: Zombies' Judgements and Epistemic Contact

A popular way to argue against the conceivability of zombies is by focusing on zombies' judgements of phenomenal consciousness. As previously mentioned, although zombies do not have any phenomenal consciousness, they are still capable of talking about them.[24] Some of them might say things like "I can see red" or "Phenomenal consciousness is mysterious." The zombie twin of David Chalmers would be writing a whole book on consciousness and arguing how the existence of phenomenal consciousness is incompatible with physicalism. The question now becomes how we should make sense of zombies' judgements on consciousness: can they be true or false? Do they have meanings? Are these judgements really about consciousness?

Some philosophers consider these judgements as the source of incoherence in the zombie idea and use them to argue that zombies are inconceivable. Thomas (1998) argues that we fail to make sense of zombies' phenomenal judgements if zombies were conceivable. Zombies' judgements can neither be true, false, nor meaningless. Balog (1999) argues that if zombies were conceivable, the zombie twin of Chalmers could have made the exact same zombie argument, which concludes that physicalism is false in the zombie world. However, physicalism is true in the zombie world, so there must be something wrong with the argument. Clark (2000) argues that if our reports of conscious experiences are causally related to consciousness itself in any way, then full-blown zombies that produce phenomenal judgements are inconceivable. Perry (2001) argues a similar point by pointing out that for zombies to make any phenomenal judgement, epiphenomenalism must be pre-supposed, which makes the zombie argument susceptible to objections to epiphenomenalism. Clark and Perry's arguments share some similarities with David Lewis's (1966) argument, which all ultimately lead to type-A materialism. Kirk (2005, 2008) also proposes a similar objection that focuses on epistemic contact. Kirk argues that the conceivability of zombies entails the conceivability of the e-qualia

[24] The "talking about them" here should be understood in the deflationary sense. Zombies are capable of uttering the word "consciousness," but these words might not have meanings.

story (a kind of epiphenomenalism), but the e-qualia story is inconceivable, which means zombies are also inconceivable. If consciousness is not physical itself and is not capable of generating any effect on any physical events, how can we know about and refer to them?

In response to objections related to epiphenomenalism (such as the ones from Perry and Kirk), Chalmers (2004a) points out that the conceivability of zombies is compatible with non-epiphenomenalist views, where consciousness plays a causal role. Russellian monism is one such view. Even if consciousness is not epiphenomenal, we can still conceive of a world that is identical to ours in all dispositional/extrinsic aspects but lacks consciousness (or is phenomenally different). Such a world will still verify the primary conceivability of zombies.

Chalmers mentioned the problem of phenomenal judgement in his original 1996 book and later touched on the issue again in his 2003b paper. First of all, Chalmers rejects the causal theory of knowledge and reference. Second, he introduces the idea of acquaintance – a special relationship that exists between a conscious experience and the experiencer. Third, he argues that acquaintances with conscious experiences are essential for forming genuine knowledge about consciousness and meaningful phenomenal judgements. So, according to Chalmers, there is a difference between zombies' phenomenal judgements and our phenomenal judgements. The difference is that we are acquainted with conscious experiences and zombies are not, and this explains why our phenomenal judgements are meaningful and zombies' are not. It is a similar case with phenomenal concepts. Zombies share the same brain states as us, which means they will have brain states that seem to represent the concept of *consciousness*. However, since zombies are not acquainted with conscious experiences, they only possess the deflated phenomenal concepts, while we possess the full-blown phenomenal concepts.

Chalmers gave a more detailed clarification of his view in his 1996 and 2003b work mentioned earlier. At first glance, the idea of acquaintance might seem a bit forced, but I think it is at least coherent and plausible. I think the problem with zombies' phenomenal judgements is not rooted in the conceivability of zombies but in the theory of meaning and reference. Even in our world, many people tend to attach meanings to uttered sounds and written words, as if the sound (or air vibration) and words (or ink strokes) themselves have meanings. However, we can easily design a robot or software that generates the English word "apple," either verbally or visually, but we probably do not think the word itself carries meaning as it is just a simple output generated by a computer. Nonetheless, when we hear or read the generated word, we seem to automatically attach a meaning to it. I think a plausible explanation is that the meaning of any physical event depends on a conscious interpreter. It is we, as conscious beings, who grant meaning to physical states. So, I think zombies' judgements are simply meaningless when

understood as objective physical events. However, we can still consider them as true or false when they are interpreted from our perspective since we do possess full-blown phenomenal concepts. I do not have the space to develop my view on reference in full detail here, but the bottom line is that I do not think zombies' phenomenal judgements cause any genuine trouble for the conceivability of zombies.

2.1.5 Objection: Other Objections Concerning the Conceivability of Zombies

There are other objections that do not neatly fit into the above categories but nonetheless concern the conceivability of zombies. For example, Marcus (2004) argues that zombies are not positively conceivable because we need to conceive of the absence of consciousness. Chalmers (2009) points out that this is a misinterpretation of positive/negative conceivability, and that there is nothing wrong with positively conceiving of the absence of things. The reply Chalmers gives here might be too brief to be convincing. Personally, I think we do not even need to appeal to positive conceivability, as we can simply say that the presence of physical states (at least in the t-physical sense) is compatible with the absence of phenomenal states, so zombies are negatively conceivable.

Frankish (2007) poses an interesting objection by generating a mirror argument to the zombie argument called "the anti-zombie argument." Anti-zombies are mere physical duplicates that have consciousness. If we replace "zombie" with "anti-zombie" in the original zombie argument, it leads to physicalism, which means the conclusions of the zombie argument and the anti-zombie argument contradict each other. Obviously, at least one of the arguments is unsound. I think that proponents of the zombie argument should deny that anti-zombies are conceivable. Mere dispositional/extrinsic (t-physical) information alone does not entail the relevant phenomenal information we want. As Nagel (1974) points out, knowing no amount of dispositional/extrinsic (t-physical) information tells us about what it is like to be a bat. If we adopt the o-physical conception, o-anti-zombies are also primarily inconceivable, as primarily conceiving of o-anti-zombies is the same as conceiving of t-anti-zombies.[25] O-anti-zombies might be secondarily conceivable and hence secondarily possible, but even so, its secondary possibility poses no threat to our argument as the two-dimensional argument is compatible with Russellian monism (or o-physicalism).

[25] This is due to the primary intensions of both o-physical and t-physical concepts involving dispositional concepts instead of categorical concepts.

Another notable objection comes from Braddon-Mitchell (2003). Braddon-Mitchell interestingly claims that phenomenal concepts have conditional structures, which pick out some non-physical states (spooky states) if the intrinsic nature of experience is non-physical, but pick out whatever physical states play such-and-such roles if the intrinsic nature of experience is physical. In short, phenomenal concepts are fixed to pick out non-physical qualia if dualism is true in the actual world and fixed to pick out certain physical states if physicalism is true in the actual world. Once we discover that physicalism is true, if it is true at all, zombies will no longer be conceivable. But since no one is absolutely certain whether physicalism is true, zombies will always seem conceivable, even if dualism is actually false. Chalmers (2009) responds to Braddon-Mitchell's objection by first pointing out that it is conceivability and not meta-conceivability that matters. Of course, Braddon-Mitchell argues that zombies are merely meta-conceivable without being really conceivable. I think that although we might not be absolutely certain about the falsity of physicalism, we can still have good reasons to support the conceivability of zombies. We, as humans, are rarely totally certain about things, but such uncertainty should not suggest the inconceivability of them. Chalmers also defends the conceivability of zombies by rejecting the conditional analysis. I think this is correct. Following my intuitions, it seems like our phenomenal concepts, unlike most other concepts such as the concept of *water*, pick out their references directly, which means we have access to the intrinsic nature of consciousness via our phenomenal concepts due to their non-conditional structures. Of course, some people might indeed possess a conditional concept of *consciousness* as Braddon-Mitchell proposes. The motivation behind such a conditional concept might be to guarantee the existence of consciousness regardless of whether dualism or physicalism is true in our world. Most people are certain about their consciousness, no matter how the world turns out. However, I think the directly-referring phenomenal concepts are capable of doing the same. I am certain that I am conscious because I have direct access to my consciousness through my phenomenal concepts. Both dualism and some form of physicalism can still be true, depending on whether consciousness (the referent of my phenomenal concepts) is also getting picked out by my physical concepts, assuming that physical concepts are referring indirectly (or opaque, to use Goff's terminology). So, there is no need for the conditional concept of *consciousness*, and even if some people do possess such a concept, it is not the concept in which we are interested in this thesis.

2.2 The Link between Conceivability and Primary Possibility

2.2.1 Reasons for Supporting the Conceivability-Possibility Thesis

The premise 2 of the argument is basically the conceivability-possibility thesis (CP). The truth of CP guarantees the truth of premise 2. Depending on whether positive or negative conceivability is used, two CP theses can be generated: (1) CP+, which claims ideal primary positive conceivability entails primary possibility; and (2) CP-, which claims ideal primary negative conceivability entails primary possibility. Chalmers (2002) is pretty certain that CP+ is true and relatively confident that CP- is also true. CP- also entails CP+. This is because something must be coherent if it can be coherently imagined. If we can somehow show that CP- is true, CP+ comes for free.

Why should we support CP-? Chalmers points out that the main reason is the lack of counterexamples. For either CP to fail, we need to find a counterexample where something is ideally primarily conceivable without being primarily possible. Chalmers (1996, 1999, 2002, 2009) calls such a counterexample a "strong necessity" or a "strong metaphysical necessity." I have briefly discussed strong necessities in Chapter 1.4, and I will clarify the idea in greater detail shortly. As we will see, there is no convincing example of strong necessities. Chalmers also calls the physicalist view that supports the existence of strong necessities "type-B materialism" (see Chalmers, 2003a). Type-B materialists deny the ontological gap but not the epistemic gap between the physical and the mental, which also means they consider zombies to be ideally conceivable but not metaphysically possible. It should be noted that if we interpret metaphysical possibility as the secondary possibility and physical as o-physical, the type-B materialism here resembles Russellian monism, which the zombie argument is not trying to defeat. However, most type-B materialists adopt the t-physical conception by default and consider zombies to be impossible even in the primary sense. So, it can be said that most type-B materialists deny the CP thesis and support some kind of strong necessities.

Is there any other independent reason to believe in CP? Chalmers (2002, 2009) thinks that our modal concepts are rooted in the rational domain. The purposes of these concepts are closely tied to the rational and psychological. Even if not all conceivable worlds are metaphysically possible, there should still be a modal concept that captures the space of ideally rationally conceivable worlds (or scenarios). This space of conceivable worlds constitutes the broadest sense of possibility. So, conceivability at least entails this broadest kind of possibility.

I agree with Chalmers here, but I think there is a simpler reason. Let us focus on negative conceivability. A scenario (or event or thing) is ideally negatively conceivable if and

only if there is no incoherence in the scenario, such that the scenario cannot be ruled out *a priori*. When we try to understand logical possibility, we come up with something similar. A scenario is logically possible if it is true in at least one logically possible world, and logically necessary if it is true in all logically possible worlds. But what makes a world logically possible? Well, one way to understand it is to say that a world is logically possible if and only if there is nothing logically incoherent about such a world. So, the space of logically possible worlds includes all the total ways a world can be, as long as there is no incoherence. The concepts of *conceivability* and *possibility* are *a priori* connected. Some might worry about the circularity issue: conceivability entails possibility because conceivability is defined using possibility. But this is not the case. Both the concept of *conceivability* and the concept of *possibility* are based on the idea of coherence. What are the differences between the two concepts then? The answer is that conceivability is subject-relative, but possibility is not. The attempt to negatively conceive of something is to think about whether a scenario is coherent. The harder we think, the closer we get to ideal conceivability. Possibility, on the other hand, concerns whether a scenario is really coherent. To know whether a scenario is possible, we simply need to try to conceive of the scenario. If there is an ideal rational thinker, such a thinker should find all logically possible scenarios conceivable and find all logically impossible scenarios inconceivable. When these scenarios are conceived as actual, primary possibility is involved. Since the primary intension does not rely on any *a posteriori* information to determine the extension in another possible world, we know *a priori* whether a sentence or proposition is coherent in any possible world when the primary intension is involved. Therefore, the link between ideal primary conceivability and primary possibility should be obvious.

2.2.2 Objection: Counterexamples and Strong Necessities

As mentioned previously, to deny CP, we simply need to find a counterexample that involves a scenario or the negation of a scenario which is ideally primarily conceivable but not primarily possible. Cases generated in the above way are called “strong necessities.” The existence of strong necessities is incompatible with the CP thesis.

To further clarify the idea of strong necessity, let us suppose that E is a strong necessity, where E is a proposition about a scenario or event. The negation of E has to be ideally primarily conceivable, which means at the very least that there is no *a priori* incoherence in the negation of E. Then, the negation of E must be false in every possible world (considered as actual), which means E must be true in every possible world (considered as actual). However, if the

negation of E is ideally primarily conceivable, it means there is a logically coherent conceived situation (considered as actual) where the negation of E is true, but somehow this situation does not constitute a possible world. So, proponents must support a kind of modal dualism, where the space of (metaphysical) possible worlds is outstripped by the space of coherent conceivable situations (see Chalmers, 2002). Finding a case of strong necessity is to discover a world that falls into the latter space but not the former space. If strong necessities do not exist, the two spaces above are identical. Chalmers (1996, 1999, 2002, 2009) has previously argued for a lack of strong necessities in great detail. I will not attempt to make any further contribution here, but I think I should at least mention a few well-known candidates for strong necessities and explain why they fail.

Let me start with the most famous case: the Kripkean *a posteriori* necessity. Consider the case "water is H_2O" again. The negation of the statement is "water is not H_2O." There is nothing *a priori* incoherent about the negation, which means it is at least negatively conceivable, but Kripke claims that it is nonetheless impossible. This is because the concept of *water* always picks out the actual local watery stuff, and the actual local watery stuff is H_2O (known to us via *a posteriori* investigations). After taking into account the *a posteriori* information that the actual local water stuff is H_2O, "water is not H_2O" is no longer coherent since "H_2O is not H_2O" is not coherent. So, one might use the case of Kripkean necessity to argue that certain scenarios or events might be *a priori* conceivable but nonetheless *a posteriori* impossible. To reply, we simply need to apply two-dimensional semantics. According to two-dimensional semantics, the concept of *water* has a hidden indexical structure whose extension in a possible world depends on whether the possible world is actual. This indexical structure means the concept has distinct primary and secondary intensions. Suppose there is a possible world where the local watery stuff is XYZ but not H_2O. When the world is considered as counterfactual, the XYZs are not picked out by the concept of *water*. However, they will be picked out by the concept of *water* if the world is considered as actual.[26] So, although there is no possible world where the secondary intension of "water is not H_2O" is true, there are possible worlds where the primary intension of it is true. The Kripkean necessity does not constitute a case of strong necessity, and Kripkean cases only show that sometimes conceivable scenarios can be primarily possible without being secondarily possible.[27]

[26] Equally, we can say that the XYZs in such a world are picked out by the primary intension of the concept but not by the secondary intension.

[27] I intend to use "Kripkean necessity" to mean the *a posteriori* necessity involved in Kripke's examples here. Some philosophers might simply use the phrase "Kripean necessity" to mean

Another well-known candidate for strong necessities concerns unsolved mathematical hypotheses. It has been widely agreed that mathematical hypotheses such as Goldbach's conjecture are either necessarily true or necessarily false, depending on whether they are true or false in the actual world. However, we seem to be able to conceive of both the truth and falsehood of Goldbach's conjecture; if so, it constitutes a case of strong necessity. This is clearly not true, and there are a few things we can say about it. If Goldbach's conjecture is *a priori* true in our world, then the negation of it should not be ideally conceivable, as the truth of the conjecture entails incoherences in its negation. On the other hand, if the conjecture is *a priori* false, then the conjecture itself should not be ideally conceivable. Let us just assume that the conjecture is *a priori* true for the sake of our discussion. Any ideal thinker will find the negation of the conjecture inconceivable and therefore will not conceive of an impossible world. Of course, we can certainly conceive of a world where the mathematicians think that Goldbach's conjecture is false and hence announce its falsehood. However, this is not a world where the conjecture is false, but simply a world where the mathematicians make the mistake of thinking that the conjecture is false, and such a world is certainly possible (see Chalmers, 2002). Furthermore, our apparent conceivability of its negation can be attributed to a sort of meta-conceivability. The negation of the conjecture might seem conceivable due to our ignorance. We might even consider it possible if we adopt the subjective conception of possibility. But this is only because we fail to notice the incoherence. In this case, we can at most say that the negation of the conjecture is *prima facie* conceivable, but since it is not ideal conceivability, the case does not count as a strong necessity.

Some candidates of strong necessities are also brought up in dedicated objections to the zombie argument. For example, Shoemaker (1999) considers laws of nature to be metaphysically necessary and uses this idea to argue against the CP thesis. In response, Chalmers (1999) first points out that there is intuitively nothing incoherent about worlds with different laws of nature. We can easily conceive of a world with a different speed of light or a different gravitational constant. If our modal concepts are used to capture our counterfactual thinking, there must be at least a broader space of possible worlds in which some of the worlds have different laws of nature. Yablo (1999), on the other hand, argues that both the existence of a necessarily existing god and the lack of a necessarily existing god are conceivable, but

---

strong necessity. If one decides to conceptualize "Kripkean necessity" in this way, I would say that the typical Kripkean examples can be explained without Kripkean necessity but with two-dimensional semantics instead. The key idea here is that strong necessity is not required to explain the Kripkean examples.

they cannot be both possible. In response, Chalmers (2009) argues that a necessarily existing god is not conceivable. The double modality here is tricky to deal with, but we can easily restrict the CP thesis to the distribution of nonmodal properties within worlds. The zombie argument is perfectly compatible with the restricted CP thesis.

Besides the four candidates mentioned above, Chalmers (2009) has also analysed plenty of others, such as cases with demonstratives, ordinary macroscopic truths, inscrutable truths, disquotational truths, and so on. I will not get into the details here. The conclusion Chalmers draws is that there is no clear example of strong necessities. The idea of strong necessities seems forced and inexplicable, and we have no good reason to believe in the existence of strong necessities.

2.2.3 Objection: The Phenomenal Concept Strategy

Instead of finding specific examples of strong necessities, opponents of the zombie argument might argue that the phenomenal-physical identity itself is a unique case of strong necessity, such that the CP thesis only fails when the mind-body problem is involved. Philosophers who adopt this idea usually argue that zombies are only conceivable due to certain special features of phenomenal concepts. Stoljar (2005) calls this "the phenomenal concept strategy." Proponents of the strategy agree with Levine (1983) that there is an explanatory gap, but only consider this gap to be epistemic and not ontological. And since this gap is only epistemic, zombies are impossible, despite them being conceivable. For this strategy to work, there must be some features of phenomenal concepts that somehow explain away the conceivability of zombies. Proponents of the strategy argue that our dualist intuitions should be well expected once we discover the unique features or characteristics of phenomenal concepts and that these intuitions are compatible with physicalist accounts of phenomenal consciousness.

There are several different versions of the strategy. Loar (1990) argues that phenomenal concepts are recognitional concepts that refer directly but nonetheless are conceptually independent of other directly referring physical concepts. He thinks that these two sorts of directly referring concepts can still refer to the same thing despite the lack of conceptual connections between them, as they have different modes of presentation. Hill and McLaughlin (1999) argue that phenomenal concepts and physical concepts belong to distinct faculties of the mind, and most dualist intuitions are purely due to the simultaneous use of these distinct faculties. As a result, we will find zombies conceivable and the explanatory gap intuitive, even though phenomenal concepts and physical concepts co-refer. Papineau (2002, 2006) argues

that phenomenal concepts are in the form of a "quotational-indexical" structure – the experience: [Blank], where the "blank" is filled with an experience. Again, he thinks that when phenomenal concepts are understood this way, there should be no surprise that an explanatory gap will arise. However, this gap in his view is merely epistemic. There are other varieties of the strategy as well, but they usually do not differ from the above ones in significant ways. The core idea of the strategy is that dualist intuitions can be explained away by simply analysing our psychological processes related to these concepts. Typically, proponents of the strategy support a form of *a posteriori* physicalism (type-B materialism).

Not everyone is content with the phenomenal concept strategy. Chalmers and Stoljar have both provided some compelling objections to the strategy. In his 2006a paper *Phenomenal Concept and the Explanatory Gap*, Chalmers provided a master argument for the failure of the strategy. The argument goes roughly as follows: First, suppose C is whatever thesis that explains our epistemic situation with regard to the explanatory gap. For C to be supportive of physicalism, C must be true and physically explicable. Second, since proponents of the strategy still consider zombies to be conceivable, we can ask the question of whether it is conceivable that zombies lack the psychological features attributed by C. Third, if it is conceivable that zombies lack the features attributed by C, then C is not physically explicable since zombies are physically identical to us. Fourth, if it is inconceivable that zombies lack the features attributed by C, then C fails to explain our epistemic situation, as zombies do not share the same epistemic situation with regard to the explanatory gap. Therefore, a dilemma is posed for the proponents of the strategy: either C is physically inexplicable or C fails to explain our epistemic situation. Both options lead to the failure of the strategy.

Stoljar (2005), on the other hand, argues that the phenomenal concept strategy only tells us that the psycho-physical conditional is not *a priori* synthesizable without telling us whether it is *a priori*.[28] Without demonstrating that the psycho-physical conditional is *a posteriori* (or not *a priori*), the strategy fails to provide support for *a posteriori* physicalism. As an alternative, Stoljar suggests *the missing concept strategy*, saying that crucial concepts are missing in our conception of the physical. Once we are able to discover these concepts, the conceivability argument will no longer work. Following Chalmers's and Stoljar's replies, we

[28] Stoljar (2005) thinks that a conditional is *a priori* synthesizable if an agent is in a position to know the truth of the conditional once they possess the concepts required to understand the antecedent of the conditional. So, a conditional might be *a priori* without being *a priori* synthesizable if the truth of the conditional is only known by an agent when the agent possesses not only the concepts involved in the antecedent but also those involved in the consequent.

can say that, at the very least, plenty of work is still left to be done before we can conclude that the phenomenal concept strategy poses any genuine threat to the zombie argument .

2.3 The Link between Secondary Impossibility and Russellian Monism

2.3.1 Reasons for Supporting the Entailment from Secondary Impossibility of Zombies to Russellian Monism

If zombies are not even primarily possible, *a priori* physicalism is true, since the relevant phenomenal facts are *a priori* entailed by the dispositional/extrinsic (t-physical) facts. If zombies are primarily possible, we are then left with the options of considering zombies as secondarily possible or secondarily impossible. According to two-dimensional semantics, we do not know which option is correct (and might never know), since we have no idea what intrinsic properties the physical concepts are picking out. Therefore, premise 3 and 4 of the argument each deals with one of the two options.

Premise 3 says that the secondary impossibility of zombies leads to Russellian monism. Why is that the case? To achieve the secondary impossibility of zombies, we want phenomenal concepts and physical concepts to co-refer. Our physical concepts are closely related to dispositional/extrinsic properties, but we already know that dispositional/extrinsic properties do not entail phenomenal properties, and that they are certainly not identical to each other. It is just like the property of being watery is different from the property of being H_2O; we can easily separate them in a possible world. Furthermore, most philosophers think that phenomenal concepts refer directly and pick out the essence (or intrinsic nature) of the referents. So, for phenomenal concepts and physical concepts to co-refer, the only way is to allow physical concepts to have distinct primary and secondary intension. To use Stoljar's (2001a) terminology, we can say that o-physical concepts must be used instead of t-physical concepts. This is similar to the case of water and H_2O. The concept of *water* has distinct primary and secondary intensions: the concept intends to pick out some hidden properties using the watery properties as a guide. So, even though we can easily separate the watery property from the H_2O property, as long as the hidden property is identical to the H_2O property, it is secondarily impossible that water is not H_2O. Similarly, o-physical concepts intend to pick out some hidden intrinsic properties about our world using the dispositional/extrinsic properties as a guide. In conclusion, the only way for zombies to be secondarily impossible, assuming that phenomenal concepts refer directly, is for physical concepts to have distinct primary and secondary intensions.

For many philosophers, the co-reference of physical concepts and phenomenal concepts alone is enough to defuse any anti-physicalist objections. But we need to be more careful here. The so-called physical concepts used here are quite different from the common physical concepts many people are familiar with. These concepts pick out hidden intrinsic properties of our world that physics has no direct access to. So, instead of physicalism, it is much more appropriate to say that the secondary impossibility of zombies leads to Russellian monism. I have discussed Russellian monism in more detail in Chapter 1.6. In short, it is the view that there is one kind of fundamental intrinsic properties (quiddities) that is picked out by both our phenomenal and o-physical concepts. Depending on what the nature of quiddities is or even just how we like to call these quiddities, Russellian monism can resemble a kind of idealism, neutral monism, or physicalism. Because of this special feature, Chalmers does not fully rule out physicalism in the two-dimensional zombie argument. What he does rule out is structural physicalism (or t-physicalism), where only the dispositional/extrinsic properties are considered as physical properties.

The differences between Russellian monism and other kinds of physicalism are subtle but important. Russellian monism can be seen as a special kind of type-B materialism since o-zombies are primarily conceivable but secondarily impossible, or a special kind of type-A materialism since the complete knowledge of o-physical facts *a priori* entails all phenomenal facts. However, I do not think most type-B materialists and type-A materialists will find Russellian monism appealing. Most type-B materialists use t-physical concepts by default and do not consider there to be any possible worlds where zombies exist. Also, according to Russellian monism, o-physical facts are unknowable unless they are also the phenomenal facts, which most type-A materialists would disagree with.

Overall, I think it is safest not to consider Russellian monism a kind of physicalism, even though the dispute here might be purely verbal. The secondary possibility of zombies suggests the same kind of properties are getting picked out by both phenomenal concepts and o-physical concepts. However, there is an asymmetry here between the phenomenal concepts and the o-physical concepts. Phenomenal concepts pick out their referents directly, which means the intrinsic nature (or essence) of the referents is transparent to the concepts. We know about the intrinsic nature of phenomenal consciousness just by being acquainted with it. On the other hand, o-physical concepts pick out their referents indirectly, which means the intrinsic nature of the referents is hidden from the concepts. Suppose that all o-physical concepts co-refer with some phenomenal concepts; phenomenal concepts have the advantage of telling us

more about the nature of the referents.[29] Moreover, since phenomenal concepts directly pick out the phenomenal properties, we can even consider phenomenal properties to be the sole kind of fundamental intrinsic properties of our world. O-physical concepts simply pick out the phenomenal properties via the dispositional/extrinsic properties. Following the above analysis, it seems that if we have to choose between idealism and physicalism as the consequence of the secondary impossibility of zombies, idealism is the obvious choice. I am not intending to provide a thorough defence of the idea here, but at least at first glance, I do not find any good reason to call the resulting view here "physicalism." Stoljar (2001a) also mentions the possible case where only some quiddities are phenomenal properties. This will likely rule out Russellian idealism and panpsychism. But instead of physicalism, the resulting view in this case is more appropriately to be called "Russellian pluralism" (Chalmers, 2015a). I will talk more about this shortly. In any case, it is safe to say that the secondary impossibility of zombies leads to some kind of Russellian thesis.

2.3.2 Objections Concerning the Link between the Secondary Impossibility of Zombies and Russellian Monism

Few, if there are any, objections focus on the link between the secondary impossibility of zombies and Russellian monism. However, many physicalists presuppose the link between the impossibility of zombies and physicalism. Many of the objections, especially those that promote *a posteriori* physicalism, consider that the co-reference of phenomenal concepts and physical concepts automatically leads to physicalism. Very often, the distinction between the primary possibility and the secondary possibility, and the distinction between t-physical concepts and o-physical concepts are ignored. It is important to recognise that once two-dimensional semantics is taken into account, the secondary impossibility (or *a posteriori* impossibility) of o-zombies (zombies that are identical to us in all o-physical aspects) only leads to a very specific view, which is best to be considered as a kind of Russellian monism. Although physicalism cannot be completely ruled out as long as there is enough room to adjust the meaning of the term "physical," structural physicalism is definitely out of the picture. So, it is certainly inappropriate to say that the secondary impossibility of zombies leads to physicalism.

Most physicalists do make the distinction between metaphysical possibility and logical possibility, but as discussed in Chapter 1.4, given that modal monism is true, the only

[29] Thanks to David Braddon-Mitchell for the discussion here.

appropriate way to interpret metaphysical possibility is to understand it as the secondary possibility. This is also why the term "metaphysical possibility" is not used in the argument. The primary/secondary distinction better reflects our modal intuitions.

Daniel Stoljar is one of the few major philosophers who thinks the secondary impossibility of zombies leads to some kind of physicalism, and since he is the one who introduces the distinction between t-physical concepts and o-physical concepts, his idea deserves a more rigorous examination here. Stoljar (2001a, 2001b) recognises that it is the t-zombies that are ideally conceivable, not the o-zombies.[30] So, we do not know whether o-zombies are possible, and if o-zombies are impossible, o-physicalism is true. That is, any minimal o-physical duplicate of our world will be phenomenally identical to our world. This is compatible with the two-dimensional argument. The real dispute here is whether o-physicalism should be considered as a kind of physicalism. Although I have been using the terms "o-physical" and "o-physicalism" throughout this thesis, I do not think that the resulting view here should be considered as a kind of physicalism. There are two things to say about it.

First, the dispute here seems to be wholly verbal. A dispute is wholly verbal when disagreements about a statement are completely due to disagreements about the meaning of a specific term (see Chalmers, 2011). The term of interest here is "physicalism." We seem to reach an agreement on the resulting ontological view but disagree on whether such a view should be called "physicalism." Usually, verbal disputes are not substantial, which means once we agree upon the meaning of the disputed term, no more disagreement will remain. So, it might seem like as long as physicalists are willing to adjust the meaning of "physicalism" to accommodate the presented view here, we should allow it to be called "physicalism." However, it is not that simple. Choices of words often have significant pragmatic influences. Saying "physicalism is true" can be misleading to people who are not familiar with the details of the discussion. As mentioned previously, the most commonly used physical concepts are the t-physical concepts, but we have also shown that t-physicalism is definitely false. So, unless the t-physical/o-physical distinction has been widely known and accepted by the philosophical community, it is best to avoid calling the view here "physicalism" to prevent misunderstandings.

Second, as mentioned in the last section, if all o-physical concepts co-refer with some phenomenal concepts, the resulting view should be better considered a kind of idealism or

[30] Stoljar prefers using the distinction of strong vs. weak conceivability, but I do not think the ways of analysing conceivability matter too much here. We can simply focus on the secondary impossibility of zombies.

panpsychism rather than physicalism, since phenomenal concepts allow direct access to the intrinsic nature of the referents. However, Stoljar (2001a) points out that it might be possible that only some, and not all, quiddities have phenomenal qualities. If this is the case, all quiddities are being picked out by o-physical concepts, but only some quiddities are also being picked out by phenomenal concepts. Therefore, the resulting view fails to constitute a kind of idealism or panpsychism, and the quiddities should be better considered a special kind of physical properties. In response to Stoljar, I think that if it is the case that some quiddities end up being non-phenomenal properties, we will end up with a kind of dualism instead of physicalism. We will need an explanation of why some quiddities have phenomenal qualities while others do not. Physicalism simply does not do the job here. At most, we can argue that the view here should be called "Russellian pluralism" instead of "Russellian monism" (Chalmers, 2015a). In any way, I do not think the view here should be considered as a kind of physicalism. Furthermore, sometimes it is thought that Russellian monism is capable of becoming a version of dualism, given certain assumptions (Alter & Nagasawa, 2012; Chalmers, 2015a). It certainly sounds contradictory at first, but this is indeed how people have been using the term "Russellian monism." I will not get into the details here. The bottom line is that we can say that structural physicalism is definitely false, and the secondary impossibility of zombies leads to some kind of Russellian thesis.

2.4 The Link between Secondary Possibility and Physicalism

2.4.1 Reasons for Supporting the Entailment from the Secondary Possibility of Zombies to the Denial of Physicalism

It is widely accepted among philosophers, including physicalists, that physicalism entails the impossibility of zombies. This is especially true with the formulation of physicalism we are dealing with here (the formulation by Jackson). Of course, physicalism can be formulated as a modal thesis or a non-modal thesis, but even the non-modal formulations have modal consequences (Stoljar, 2022). As mentioned in Chapter 1.6, physicalism is the thesis that everything either is physical or supervenes on the physical. It follows that fixing the physical facts automatically fixes all the facts about our world, including the phenomenal facts. So, any minimal physical duplicate of our world must be phenomenally identical to our world. The secondary possibility of zombies contradicts this, unless we also consider our world to be phenomenally empty, which leads us back to the discussion about eliminativism.

For zombies to be secondarily possible, the referents of at least some phenomenal concepts must not be picked out by physical concepts. If some phenomenal concepts fail to co-refer with physical concepts, when we create a minimal physical duplicate of our world, some phenomenal properties will be left out, which means such a minimal physical duplicate of our world will not be a duplicate *simpliciter*. The primary possibility of zombies requires phenomenal concepts not to co-refer with dispositional/extrinsic (or t-physical) concepts. The secondary possibility of zombies requires phenomenal concepts not to co-refer with o-physical concepts. If not even o-physical concepts are capable of picking out the phenomenal properties, then phenomenal properties are over and above the physical in every sense. So, the primary possibility of zombies tells us that structural physicalism (or t-physicalism) is false, while the secondary possibility of zombies rules out all physicalism at once.

2.4.2 Objections Concerning the Link between the Secondary Possibility of Zombies and Physicalism

As it is almost a consensus that the metaphysical possibility of zombies leads to the denial of physicalism, not many objections can be found. Of course, one might argue that metaphysical possibility is not the same as secondary possibility, and the argument does not rule out the metaphysical possibility of zombies. However, in Chapter 1.4, I have mentioned that the only plausible interpretation of metaphysical possibility appears to be interpreting it as the secondary possibility. Still, some philosophers might insist on the existence of a distinct kind of metaphysical modality. As discussed earlier, for such a metaphysical modality to exist, we will need to specify a distinct space of metaphysically possible worlds where every world involved shares the same laws of metaphysics as ours, which leads to modal dualism. It is unclear what these laws of metaphysics are and what explanatory power this space of metaphysically possible worlds has. At first glance, there is no use in proposing such laws and a space. Still, arguing against the existence of metaphysical laws and modal dualism will be out of the scope of this thesis. So, I will be focusing on the primary/secondary distinction in this thesis instead.

By interpreting metaphysical possibility as the secondary possibility, the only relevant objection I have found comes from Piccinini (2017). Piccinini argues that the metaphysical possibility of zombie worlds does not lead to the failure of physicalism as long as the zombie worlds are not accessible from our world. He considers the physicalism that Jackson (1998) adopts inadequate. Instead, he promotes proper physicalism, which can be defeated only if zombies exist in a possible world that shares the same metaphysical laws with our world, as

only possible worlds with identical metaphysical laws are accessible from our world in the relevant sense. So, the conceivability begs the question by assuming that phenomenal physicalism is false in our world. If phenomenal physicalism is true in our world, zombie worlds (if possible at all) will therefore count as possible worlds with different metaphysical laws, which means they are not accessible from our world. In a direct reply to this objection, Prelevic (2017) points out that Piccinini's objection assumes modal dualism, which we have shown earlier in Chapter 1.4 that should be ruled out. Chalmers's modal monism abandons the idea of metaphysical laws, which also means there is no distinct space of metaphysical possible worlds. The distinction between logical possibility and the so-called metaphysical possibility only exists on a semantic level. If so, we only need the space of logically possible worlds to explain the so-called metaphysical modality, and all logically possible worlds are accessible from one another.

Chapter 3: Replies to Recent Objections to the Argument

In this chapter, I will reply to four objections to the zombie argument from three of the more recent papers in greater detail. The first one is Phillip Goff and David Papineau's (2014) *What's Wrong with Strong Necessities?*; the second one is Daniel Stoljar's (2020) *Chalmers v Chalmers*; and the third one is Eugen Fischer and Justin Sytsma's (2021) *Zombie Intuitions*. The first two papers are more or less purely philosophical, while the third one focuses on experimental philosophy. For each paper, I will first introduce the objection presented in the paper. Then, I will mention any notable reply to the objection if there is any. Last but not least, I will provide my own analysis of the objection and explain why the objection fails.

Why these objections? There are three main reasons. First, these three papers are relatively recent and have not yet attracted a lot of attention. There are plenty of notable objections from the late 90s to the early 2000s, but most of them have already been widely discussed. I am happy to provide my own replies to them, but my replies would probably resemble ideas that have already been previously mentioned. This is why I prefer going through these older objections in Chapter 2 rather than analysing them myself here. Second, these are some of the more significant objections that have been recently proposed. All three articles here come from very different yet interesting angles to attack the zombie argument. Especially with Fischer and Sytsma's objection, rarely has anyone previously argued against the zombie argument using experimental data. Replying to these objections guarantees that good progress is being made in defending the argument. Third, these are the objections that I think I can bring down in the most convincing way. Of course, as a proponent of the two-dimensional zombie argument, I believe it can be defended against the best objections. But replying to certain objections can be a complicated matter. Many objections rely on specific background theories, which means replying to them often requires debating on these background theories. For example, whether objections that utilise zombies' phenomenal judgements will work depends heavily on how we understand meaning and reference. What might first seem to be a debate on the zombie argument can quickly turn into a debate on the theory of meaning and reference. Of course, if the zombie argument is sound, then any objections that touch on the topic of meaning and reference will be defused, given the correct or appropriate theory of meaning and reference. However, it can easily turn into a complicated and messy process that goes beyond the scope of the current thesis. So, I chose these objections with background knowledge that I am confident of dealing with. This makes it much more manageable to analyse these objections thoroughly so that I can produce powerful defences against them.

3.1 Reply to Goff and Papineau (2014)

Chalmers's two-dimensional zombie argument endorses modal monism – the view that there only exists one primitive modal space (Chalmers, 2002; 2009). Modal monism rules out the existence of strong necessities (for modal monism and strong necessities, see Chapter 1.4 and Chapter 2.2). The only metaphysical modality that can exist is the one related to the typical Kripke-style examples, which can also be fully explained using two-dimensional semantics and modal monism. In *What's Wrong with Strong Necessities?*, Goff and Papineau (2014) produced two individual arguments against Chalmers's modal monism. In both arguments, a kind of modal dualism is supported. Although these are not arguments directly targeting the zombie argument, they nonetheless target the foundation of the zombie argument which are modal monism and the lack of strong necessities. In this section, I will introduce both Goff and Papineau's arguments against modal monism, which will be followed by my responses to their arguments.

3.1.1 Objection by Philip Goff

Philip Goff proposes a kind of modal dualism called "stress-free modal dualism." To understand his view, we first need to distinguish between the four categories of concepts (Goff, 2011):

> *Transparent concepts*: concepts that reveal the whole natures of their referents.
>
> *Translucent concepts*: concepts that reveal parts of the nature of their referents.
>
> *Mildly opaque concepts*: concepts that do not reveal the essential properties but only accidental properties of their referents.
>
> *Radically opaque concepts*: concepts that reveal neither the essential nor accidental properties of their referents.

According to stress-free modal dualism, there exists a distinction between "really" conceivable and "merely" conceivable. Something is really conceivable when only transparent concepts are involved, and something is merely conceivable when non-transparent concepts are involved. With these two kinds of conceivability, two modal spaces can be created. The space of metaphysically possible worlds consists of the worlds that we can really conceive of, and the space of epistemically possible scenarios consists of the worlds we can either really or merely conceive of. To use Goff's (2014) example, "water is not H_2O" is conceivable but not metaphysically possible because *water* is not a transparent concept. The concept of *water* only reveals the accidental properties of the referents, and as a result, we can merely conceive of "water is not H_2O" even though it is not really conceivable. Since it is merely conceivable, the

scenario does not correspond to a genuine possibility. On the other hand, a million-sided object is really conceivable since the concept of *million-sided object* is transparent.

Why should we accept stress-free modal dualism rather than Chalmers's modal monism? Goff argues that his stress-free modal dualism can deal with radically opaque concepts, which Chalmers's modal monism cannot. This makes Goff's view preferable to Chalmers's. If we endorse stress-free modal dualism, strong necessities can be created by drawing an identity sign between any two distinct co-referring radically opaque concepts. Although Philip Goff himself is an anti-physicalist, his stress-free modal dualism is incompatible with the zombie argument: if strong necessities existed, zombies might be metaphysically possible without Russellian monism being true.

3.1.2 My Reply to Philip Goff

I have in total four replies to Goff's objections. My first reply is that Goff's distinction between transparent, translucent, and mildly opaque concepts seems to be compatible with Chalmers's modal monism and two-dimensional semantics (I will deal with radically opaque concepts shortly). Applying two-dimensional semantics, transparent concepts are concepts whose primary and secondary intensions are identical. It does not matter if the world is conceived of as actual or counterfactual; transparent concepts always pick out the same referents in any possible world. Non-transparent concepts, on the other hand, are concepts with distinct primary and secondary intensions. Since these concepts do not capture the total essence of their referents, the primary intension might pick out things of different natures in different worlds. This can happen when two worlds share identical accidental features without sharing the totality of essential features, or share some identical essential features without sharing the totality of essential features. Furthermore, the primary intensions of mildly opaque concepts are capable of picking out things of completely different natures in different worlds, while the primary intensions of translucent concepts always pick out things that at least overlap in some of their natures in different worlds. Two-dimensional semantics perfectly explains the distinction between transparent, translucent, and mildly opaque concepts.

My second reply concerns the idea that metaphysical possibility cannot arise when non-transparent concepts are involved. As Goff himself points out, the concept of *water* is not transparent, which means that the conceivability of water not being H_2O fails to correspond to a genuine possibility. However, this also means that the conceivability of water is H_2O does not correspond to a genuine possibility due to *water* being a non-transparent concept. I do not think this is a desirable consequence for Goff, as most philosophers, probably including Goff

himself, consider "water is H_2O" as metaphysically possible at the very least. So, something is either wrong or unclarified in Goff's framework. On the other hand, Chalmers's modal monism has no problem dealing with this case.

My third reply concerns the phrase "genuine possibility." Goff claims that "it is conceivable that water is not H_2O, but this conception does not correspond to a genuine possibility" (p. 754). However, according to two-dimensional semantics, when we conceive of water not being H_2O, there is a genuine possible world in play – the possible world where the local watery stuff is not H_2O.[31] There should not be any objection to the existence of such a possible world. The real question is whether the claim "water is not H_2O" is true in this possible world. If the concept of *water* is *the actual local watery stuff*, then the answer to the question above lies in whether the possible world is considered as actual or counterfactual. The move that Goff makes here seems to be fixing our world as the actual world and claims that there is no way for the claim "water is not H_2O" to possibly be true in any world. Even if we follow Goff's move here, there will still only be one space of possible world – the same space utilised in modal monism. Goff simply forbids the use of primary intensions when evaluating sentences, and when this approach is adopted, "water is not H_2O" is not possible even in the broadest sense. This is not the move that I endorse as I think two-dimensional semantics does a much better job of explaining our modal intuitions. However, no matter which option we adopt here, there is no path leading to modal dualism.

My fourth reply is the most important one. I agree that Chalmers's modal monism is incompatible with the existence of radically opaque concepts, but also that we have no good reason to believe in such concepts. According to Goff, these concepts reveal neither accidental nor essential properties of the referents. These concepts seem empty to me, and it is unclear how they are capable of picking out any referent. One reply might be to rely on the causal-historical theory of reference, saying that the concepts can simply pick out the things that are causally linked to the concepts in an appropriate way without relying on any extra properties of the things. But even so, these concepts do not seem to be radically opaque. Let us say one such concept is X. The concept of X picks out the things that are causally linked to the concept in an appropriate way and has no extra content besides that. Even if that is the case, X still reveals at least one accidental property of the referents, i.e., the relational property of being causally linked to the concept in an appropriate way. X is not a radically opaque concept.

[31] The local watery stuff can be understood as the things in the tap, lake, or rain that share all the apparent watery properties such as being transparent, odourless, and drinkable.

Philip Goff gave an example of a potentially radically opaque concept in one of his earlier (2011) papers:

> Suppose I meet Bob at a party and form a recognitional capacity which allows me to refer to Bob. Two months later I wake up and think, 'I wonder if I'll me[e]t that guy again', where I employ my recognitional capacity in order to think about Bob. I can't remember Bob's name, or what he looks like. The only thing I remember about Bob is that he is a guy I met at a party, but this does not uniquely identify him, as I have met lots of guys at parties. This seems to me a plausible example of a radically opaque concept. (p. 196, footnote 11)

I disagree that the concept used in the above example is radically opaque. If the concept does not uniquely identify Bob, does the concept really pick out Bob as its referent? Something sounds off in this example. Even if I provide all the information about the party (including the information about every person in the party, all the communications that occurred, how every interaction is encoded mentally by the subject, etc.) to the person that possesses such a concept (Philip Goff in this case), Bob still cannot be uniquely identified. The purpose of the concept is defeated. The intension of the concept is undetermined since the extension cannot be determined even when the complete information about the scenario is provided. Once again, a proponent of radically opaque concepts can rely on the causal-historical theory of reference and claim that the concept picks out the referent via the causal link between the concept and the referent. But as I have explained earlier, such a causal link to the concept should also be considered as an accidental property of the referent. The concept is not radically opaque, as it still reveals this accidental property.

Without any convincing examples, there is no reason to believe in the existence of radically opaque concepts, and no reason to suppose any concept as radically opaque. As I mentioned above, radically opaque concepts appear counterintuitive and empty. I fail to see how these concepts are capable of referring in any way. In the same paper, Goff also says that "if 'Cicero' and 'Tully' are both radically opaque terms, then 'Cicero is Tully' is conceivably false even though there is no genuine possibility corresponding to its falsity" (p. 755). To object to this, I shall point out that the antecedent of the conditional is false, as "Cicero" and "Tully" both reveal some accidental properties of the referents, given the reasons provided above.[32] Still, Goff might not be focusing on the meanings of the term "Cicero" and "Tully" here, but

[32] Braddon-Mitchell (2004) also provides a similar argument for why the meanings of "Cicero" and "Tully" are not exhausted by their reference.

only to use this example to demonstrate the possibility of putting an identity sign between two radically opaque terms to create strong necessities. But as I have demonstrated above, such a possibility cannot happen if no concept is radically opaque. If there is no radically opaque concept or term, it seems like Chalmers's modal monism should be preferred over Goff's stress-free modal dualism as the former has already offered an elegant and trouble-free account of our intuitions about conceivability and possibility.

3.1.3 Objection by David Papineau

The main idea behind Papineau's objection is that "metaphysical modality has nothing to do with conceivability, … [but instead] is grounded in counterfactual thinking …[which] reflects the causal structure of the world" (p. 756). To support his idea, he focuses on two claims: (1) "David Papineau's father is Owen Papineau"; and (2) "David Papineau's birthplace is Como." He thinks that the first claim about fathers is metaphysically necessary but the second claim about birthplaces is not. He then moves on to argue that conceivability is incapable of explaining the asymmetry between the modal status of the two claims. The complete details are too messy to be explained here. Some of the details will be mentioned later when I explain my reply.

Near the end of his objection, Papineau returns to the example he discussed in his (2002) book *Thinking about Consciousness* – "Cicero is Tully." The claim is a classic example of metaphysical necessities where "Cicero" and "Tully," despite being distinct concepts, necessarily co-refer. It is conceivable that "Cicero is not Tully" but not metaphysically possible. One question Papineau asks is that if there is a conceivable situation where Cicero is not Tully, then which one of them shares the same identity as the real (or actual) Cicero? Papineau does not think the modal monists can answer this question, as he thinks that modal monism does not allow real people to inhabit the non-actual worlds. Combined with the reasons given in the earlier parts of his objection, he concludes that conceivability does not provide us access to the space of metaphysical modality. What follows is that the conceivability of zombies does not lead to the metaphysical possibility of zombies.

3.1.4 My Reply to Papineau

First of all, I think David Papineau might have some misunderstandings about the details of modal monism and two-dimensional semantics. Modal monism and two-dimensional semantics do not see metaphysical modality (or secondary modality) as grounded in conceivability alone. To know whether a claim is metaphysically possible, *a posteriori*

information is needed. What modal monism and two-dimensional semantics say is that one space of logically possible worlds plus two ways of thinking about the worlds plus *a posteriori* information can fully explain metaphysical modality. There is no need for two distinct modal spaces. In a way, the difference between logical modality and metaphysical modality occurs on a semantical level – it depends on what specific things are being picked out in a given possible world by the concepts or terms. Conceivability matters here because it is closely related to how the space of logically possible worlds is constructed, which I will get into shortly.

I should also clarify again that the two-dimensional zombie argument does not rule out the metaphysical impossibility of zombies. This could happen if certain physical concepts and phenomenal concepts co-refer. If that is the case, claims like "pain is the firing of the c-fibres" might be metaphysically necessary. However, the resulting ontological view will be quite different from traditional physicalism, due to the physical concepts involved being mildly opaque (using Goff's terminology) but the phenomenal concepts involved being transparent. This means that our phenomenal concepts, not physical concepts, capture the essence of the referents. Therefore, it resembles more closely to idealism than physicalism. So, the metaphysical possibility of zombies is not a problem for the zombie argument.

Papineau thinks that metaphysical modality is grounded in counterfactual thinking, not conceivability. However, I can try to provide some plausible explanations on how counterfactual thinking and conceivability might connect. Counterfactual thinking arguably relies on the use of possible worlds, and possible worlds are based on the idea of coherence. Although there is not yet a consensus on what makes a world possible, people commonly consider logically possible worlds as worlds that contain no incoherence. Conceivability, especially negative conceivability, is also closely related to the idea of coherence. A scenario is negatively conceivable if and only if it is free of incoherence (Chalmers, 2002). The link between negative conceivability and the space of logically possible worlds should then be obvious. There should also be little surprise as to why conceivability is closely related to modality.

Now, I want to move on to the discussion on fathers and birthplaces. Papineau spends more than half of his section focusing on why conceivability is incapable of explaining the modal asymmetry between the statement on fathers and the statement on birthplaces. There are two relevant statements here:

(1) David Papineau's father is Owen Papineau.

(2) David Papineau's birthplace is Como.

Both statements are true in the actual world, but Papineau thinks that (1) is metaphysically necessary but (2) is merely metaphysically possible, as it is also metaphysically possible that David Papineau's birthplace is not Como. If I can successfully explain the asymmetry here using modal monism and two-dimensional semantics, Papineau's objection loses most of its force.

The first question is why (1) should be metaphysically necessary. This requires us to first understand the concept of *David Papineau*. The concept of *David Papineau* is most likely to be mildly opaque,[33] picking out a set of essential properties of the referent via a set of accidental properties. In footnote 10 of Goff and Papineau's (2014) paper, Papineau briefly mentions the assumption that parents are essential to their children, which means at least two of the essential properties of any personal concept (such as the concept *David Papineau*) are having X as father and Y as the mother where X and Y are the actual parents of the actual referent of the concept. Although I disagree with his conception of personal identity, let us just follow his idea for now.[34] However, it is not hard to see that these are not the only two essential properties, as David's sisters and brothers (if he has any) do not get picked out by the concept. The rest of the essential properties will depend on the specific view of personal identity. Here we can just say that the rest of the essential properties are Z, where Z is either a single or a conjunct of properties. Also, since we know that David's father is Owen Papineau via *a posteriori* information, but not the identity of his mother, we can merge the property of having Y as the mother and the property of Z and create Z*.[35] So, the essential properties that the concept of *David Papineau* picks out are the property of having Owen Papineau as the father and the property of Z*, where Z* is a conjunct of properties.

---

[33] It might also be translucent, but cannot be transparent or radically opaque. The concept will be translucent if some essential properties, such as the property of being a human being and the property of having parents, are included in the concept.

[34] We can easily create a concept of *personal identity* that does not pick out one's parents as part of the essential nature. For example, we can create a concept that picks out one's birthplace as essential. With this new concept, (1) will not be metaphysically necessary, but (2) will. Of course, such a concept might not be pragmatically useful, but it is not wrong. Papineau's objection here depends on his specific concept of *personal identity*, but as I will demonstrate later, even with his concept of choice, his argument remains problematic.

[35] If the identities of the parents are essential to the identity of the children, then the identities of the parents will depend on the identities of their parents. To locate Owen Papineau in other possible worlds will then require us to locate his grandparents. This leads to a sort of regress. To simplify the matter, I shall assume that we somehow find a way to locate the real Owen Papineau in each possible world. This allows us to focus on the concept of *David Papineau*.

What about the accidental properties? I should clarify first that different people will have different concepts of *David Papineau* that contain different sets of accidental properties. Some of the candidates might be looking and behaving like such-and-such, being in front of me at *that* time,[36] being named "David Papineau," having written the book *Thinking about consciousness*, being a British philosopher, and so on. Here, the chosen accidental properties serve as ways of identifying the referent. It is important to remember that any property chosen here, let us say the property of Q, will constitute the tautology in the form of "David Papineau is Q." This is similar to "water is the actual local watery stuff" being a tautology, where "the actual local watery stuff" is simply the content of the concept of *water*. So, if I choose the property of having written the book *Thinking about consciousness* as the sole component of my concept, my concept will pick out anyone or anything that writes that book, even if it was a robot that did it.[37] Coming up with the exact explicit analysis is, thankfully, not required in the current discussion. The real analysis of the concept will likely contain a bunch of accidental properties that are weighted differently. For the sake of simplicity, I shall just say the accidental property relevant to the concept of *David Papineau* is U, where U is a set of properties with different weights.

At this point, you might wonder why I spend so much time discussing the concept. This is because the content of the concept is what matters here. The way we construct the concept determines the modal status of claims such as (1) and (2). To recap, the concept of *David Papineau* we have constructed picks out the thing with the essential properties of having X as the father and the set of essential properties Z* via accidental properties U. Furthermore, we know that X is Owen Papineau via *a posteriori* information. According to two-dimensional semantics, the secondary intension of the concept picks out whatever has Owen Papineau as the father and also the set of properties Z* in any given possible world. The primary intension, on the other hand, picks out whatever has the properties U in any given world. Why? This is because the concept of *David Papineau* is built to pick out the set of essential properties in the actual world, not the set of accidental properties. However, the set of accidental properties acts as the tool that leads us to the essential properties. So, when a possible world is considered as actual, the same set of accidental properties might lead us to a different set of essential properties.

---

[36] This is a case of fixing the referent via ostension, see Kripke (1972).

[37] If being a person or having parents is an essential feature, then the possible world where a robot writes the book *Thinking about consciousness* will be a world where nothing is being picked out by the concept of *David Papineau*.

Let me clarify by constructing some possible worlds. I can conceive of a world where someone is a British philosopher, writes the book *Thinking about consciousness*, and is named "David Papineau," but has Philip Papineau as the father. This person does not get picked out by the secondary intension of the concept *David Papineau*, as it lacks at least one of the essential properties – the property of having Owen Papineau as the father. However, this person might still get picked out by the primary intension as long as it satisfies the set of accidental properties U. This is because when the world is considered as actual, the concept picks out the essential property of having Philip Papineau as the father, not Owen Papineau as the father. At the same time, I can conceive of a world where someone is a plumber, never writes the book *Thinking about consciousness*, and is named "Peter Papineau," but has Owen Papineau as the father and the set of essential properties Z*.[38] This person does get picked out by the secondary intension of the concept since all the essential properties are satisfied.

It should now be clear why (1) seems metaphysically necessary (or secondarily necessary). The secondary intension of the concept *David Papineau* picks out the person who has Owen Papineau as the father and Z* in every possible world. This is because having Owen Papineau as the father is one of the essential properties of the referent of the concept. The primary intension of the concept, however, does not always pick out the person that has Owen Papineau as the father, but always picks out the person that satisfies U. Therefore, "David Papineau's father is not Owen Papineau" is conceivable and primarily possible, in which the concept of *David Papineau* picks out the person that satisfies U but not the property of having Owen Papineau as the father or Z*.

What about the claim about birthplace? How should we analyse the modal status of (2)? (2) is not metaphysically necessary due to the property of being born in Como not counting as an essential property of the concept *David Papineau*. If such a property did count as an essential property to the concept, then (2) is also metaphysically necessary. What about the primary modal status then? It depends on whether birthplace is included in the set of U. If birthplace was a property in U, "David Papineau is born in Como" will be primarily necessary. Just like it is primarily necessary that "water is watery." There might be a world where H_2O does not appear watery but greasy.[39] There is no way for the residents of that world to pick out H_2O with

[38] Obviously, the set of essential properties here must not contain properties related to profession, publications, and name.

[39] The secondary intension of water will still pick out the greasy H2O as its referent, which means "water is not watery" is primarily impossible and secondarily contingent. This might be a good reason to avoid equaling primary modality to logical modality and equaling secondary modality to metaphysical modality, as we probably want to avoid saying certain things are

the concept of *water* since the concept of *water* is *the actual local watery stuff*. However, it is most likely that birthplace is not included in the concept of *David Papineau*, so "David Papineau is not born in Como" is both primarily and secondarily possible.

In short, the asymmetry in modal statuses between (1) and (2) is due to the nature of the concept *David Papineau*. Parenthood counts as a part of the essence of the referent of the concept, but birthplace does not. A single space of possible worlds (the space of logically possible worlds), plus two ways of considering the world (actually and counterfactually), plus conceptual analysis, plus *a posteriori* information give us all information about the primary and metaphysical modality (if we interpret metaphysical modality as secondary modality). There is no need to posit two distinct modal spaces. Conceivability plays its role by providing us a single space of possible worlds.

Just an extra note. I mentioned earlier that I disagree that parenthood is essential to personal identity. I can conceive of such a possible world:

> Biological technologies are so advanced in this world that people can be made in machines with any desired genetic structure selected. One of such machine-made people shares the exact identical genetic structure as the actual David Papineau but has no parents. This person is also named by the society "David Papineau."

Does this person get picked out by the secondary intension of the concept *David Papineau*? Papineau would say no since this person does not have Owen Papineau as the father, but I think there are strong intuitive appeals that this person does get picked out by the secondary intension of the concept. The disagreement here is purely conceptual. There is no real disagreement on the nature of this possible world. The only disagreement exists on whether this person gets picked out by the secondary intension of the concept, which will depend on what essential properties we want the concept to pick out. If my concept of *David Papineau* is not built to pick out the parenthood but only the genetic structure, then this person counts as the genuine counterpart of the real David Papineau, and (1) will be merely metaphysically contingent. Which concept should be used then? The answer to this question depends on which theory of personal identity that we adopt. The discussion on personal identity is way out of the scope of this paper, but no matter what theory we adopt, and no matter what concept of *David Papineau* we adopt, modal monism is capable of fully explaining our modal intuitions.

---

logically impossible but metaphysically possible. Instead, it is better to see primary modality and secondary modality as two kinds of logical possibility.

Finally, we can get to the example of "Cicero is Tully." It is conceivable that "Cicero is not Tully" with a possible world where two distinct people are picked out by the concept of *Cicero* and the concept of *Tully* accordingly. Now, Papineau asks which one of them in this world shares the same identity as the real Cicero (or Tully). The answer will be it depends. If we adopt Papineau's conception of personal identity, whoever shares the same parents and other essential properties as the actual Cicero will be the real counterpart of the actual Cicero. This might be the "Cicero" or the "Tully" or none of them in the conceived possible world. Trans-world existence, unlike what Papineau thinks, does not really pose any problems for modal monism. The only ambiguity in the answer is due to ambiguity in our theory of personal identity – we have not yet reached a consensus on what constitutes the essence of a person. Modal monism does allow real people to be in non-actual scenarios or worlds and explains modal questions that we are interested in.

3.2 Reply to Stoljar (2020)

Stoljar (2020) proposed an objection to the zombie argument in his paper *Chalmers v Chalmers* by pointing out an inconsistency between the structuralism in David Chalmers's (2012) book *Constructing the World* and the zombie argument. In the following, I will first explain Stoljar's objection and the relevant background information. Then, I will briefly mention how Chalmers himself replies to the objection. Finally, I will present my own reply to Stoljar's objection.

3.2.1 The Objection

Before I get into the explanation, I shall first lay out the argument provided by Stoljar and the relevant theses. Stoljar thinks that Chalmers supports the following three theses (p. 470):

(1) The no entailment thesis: For any phenomenal truth T*, there is no physical or topic-neutral truth T, such that T *a priori* entails T*.

(2) Chalmers's characterization of the physical: For any physical truth T, there is a truth R such that T is *a priori* equivalent to R, where R is a Ramsey sentence whose o-terms are (a) logical/mathematical (b) causal/nomic and (c) spatiotemporal.

(3) Phenomenal functionalism: For any spatiotemporal term α, a truth of the form 'there is an x such that x is α' is *a priori* equivalent to a truth of the form 'there is an x such that x is the normal cause of experience E'.

A few clarifications here. First, Stoljar claims that (1) is *a priori* equivalent to the first premise of the zombie argument (zombies are ideally conceivable). This seems *prima facie* correct since if phenomenal truths about us are entailed by physical or topic-neutral truths, then minimal physical duplicates of us will be automatically phenomenally conscious. Second, Stoljar claims that Chalmers uses (2) as the characterization of physical truths. I will not get into the details of Ramsey sentences here, but commonly, we can understand them as a way to make sense of unobservable theoretical terms using terms that we are familiar with. Basically, for some sentences containing such-and-such terms, we can find a Ramsey sentence that is *a priori* equivalent to it but contains no such-and-such term. The whole process can be seen as a process of reduction or term replacement. The method of creating Ramsey sentences about these new theoretical terms is called the method of *Ramsification* (For information on Ramsey sentences and Ramsification, see Lewis, 1970; Lewis, 2001; Chalmers, 2012, Chapter 7). Basically, (2) is saying that any physical truth that contains non-spatiotemporal physical terms such as "mass," "electron," and "H_2O" can be transformed into an equivalent truth that contains only logical/mathematical, causal nomic, and spatiotemporal terms. In the transformed truth, which is a Ramsey sentence, the non-spatiotemporal physical terms are eliminated. Finally, Stoljar claims that Chalmers adopts (3) in his *book Constructing the world*. If (3) is true, then any truth containing spatiotemporal terms can be transformed into an equivalent truth that contains no spatiotemporal terms. Instead, these spatiotemporal terms will be replaced by certain logical/mathematical, causal/nomic, and phenomenal terms. (2) and (3) are similar in the sense that (2) is capable of eliminating non-spatiotemporal physical terms and (3) is capable of eliminating spatiotemporal terms. By combining phenomenal functionalism and other ideas in Chalmers's (2012) book, Stoljar claims that Chalmers is committed to a certain structuralism (p. 474):

> (4) Structuralism: For every truth T, there is a truth B formulable in topic-neutral or phenomenal terms such that B *a priori* entails T.

Stoljar then goes on to argue that there is an inconsistency between structuralism and the zombie argument, which means one or both of them have to be given up. Before I get into the argument, I shall first explain a bit of my own intuitions on phenomenal functionalism and structuralism. Let me focus on phenomenal functionalism here. One way to demonstrate the plausibility of (3) is to investigate our own physical concepts[40]. Suppose the concept of *water*

[40] Chalmers's discussion in his (2012) book focuses on sentences and terms instead of propositions and concepts. His justification is given in Chapter 2 of his book. Here, I focus my

is *the actual local watery stuff*. What is the concept of *watery*? Well, it is intuitive to think that the concept of *watery* picks out certain physical properties. But how do we pick out these properties? The only way that makes sense to me is via our conscious experiences: our unique (visual, olfactory, tactile…) experiences associated with water. For something to be watery is just to be capable of generating water-related experiences under the right conditions. The same applies to other physical concepts such as *chair*, *mass*, and *electron*. Electrons might not be directly observable, but we propose the existence of electrons to explain a variety of observable phenomena. Hence, the concept of *electron* is likely to be based on certain phenomenal concepts as well. The most extreme case is to draw the link between spatiotemporal concepts and spatiotemporal experiences. That is, for something to be space or time is just to bring about specific spatial or temporal experiences. If one decides to take this step, then sentences containing spatiotemporal terms (which represent spatiotemporal concepts) can be transformed into equivalent sentences that contain no spatiotemporal terms but phenomenal terms (which represent phenomenal concepts). This is the case for (3). From (3), we can get structuralism by further saying that every truth is *a priori* entailed by a truth that is formulated using spatiotemporal, topic-neutral, and phenomenal terms. I am by no means properly defending phenomenal functionalism and structuralism here. However, I do consider them plausible and coherent. My explanation above only intends to provide some rough insights into their plausibility and coherence.

Let us move on to Stoljar's argument. His argument is divided into two phases. The first phase contains the following (p. 475):

> (P1) There is a truth R such that R is *a priori* equivalent to S, where R is a Ramsey sentence whose o-terms are (a) logical/mathematical, (b) causal/nomic and (c) spatiotemporal.
>
> (P2) There is a truth D such that D is *a priori* equivalent to R, where D is exactly like R except that every spatiotemporal expression α in R has been replaced with an expression of the form 'is the normal cause of experience E.'
>
> (P3) D is a phenomenal dispositional truth.
>
> (C1) S is *a priori* equivalent to a phenomenal dispositional truth, viz., D.

---

discussion on concepts, as it is a more intuitive way of understanding the matter. The difference between concepts and terms here does not affect the overall discussion.

S is an arbitrary physical truth, and the argument above is valid. (P1) follows from (2); (P2) follows from (3); (P3) follows from the definition of phenomenal dispositional truths. In terms of the idea of phenomenal dispositional truths, Stoljar gives the following clarification:

> Phenomenal dispositional truths are those that entail that something is disposed to produce a certain experience. So understood, phenomenal dispositional truths are not necessarily phenomenal truths, that is, those that entail that someone has an experience. For something can be disposed to produce an experience without ever doing so, just as a sugar cube can be soluble without ever dissolving. (2020, p. 475)

Stoljar considers (C1) to be a position Chalmers supports. The second phase of Stoljar's argument contains the following (p. 475):

> (P4) If D is *a priori* equivalent to a physical truth, D+ is *a priori* equivalent to a physical truth [where D+ is the conjunction of D and 'x causes what it normally causes'].
>
> (P5) D is *a priori* equivalent to a physical truth.
>
> (C3) D+ is *a priori* equivalent to a physical truth.
>
> (P6) D+ *a priori* entails a phenomenal truth, viz., that E occurs.
>
> (C4) There is a physical truth that *a priori* entails a phenomenal truth; that is, the no entailment thesis is false.

This argument is also valid. Since Stoljar thinks that the no entailment thesis is *a priori* equivalent to the first premise of the zombie argument, the argument presented here also serves as an objection to the zombie argument. Three options are offered by Stoljar: to give up structuralism, to give up the zombie argument, or to give up both. Stoljar promotes a kind of type-C materialism that gives up both structuralism and the zombie argument. I will not get into his view here as I do not consider the argument here to be sound. I think none of the three options needs to be chosen.

3.2.2 Chalmers's Reply

Despite being such a recent objection, Chalmers (2020) himself has already formulated a reply and published it in the same issue as Stoljar's article. The first move Chalmers makes is to deny the role of (1) in the zombie argument. The strong no entailment thesis is not necessary for the zombie argument. For the first premise of the zombie argument to work, we do not need to conceive of a world where no consciousness exists. All we need is to conceive of a world that is physically identical but phenomenally different in some respects. So, it seems like dualism can still survive even if some phenomenal truths are entailed by physical truths,

as long as not all phenomenal truths are entailed by physical truths. I agree with Chalmers's move here, and I will provide a more detailed analysis of this idea in the next section.

The major move from Chalmers, however, is to deny (3), which is phenomenal functionalism. By denying (3), (P2) from the first phase of Stoljar's argument is also denied. Put simply, Chalmers clarifies that (3) is a misunderstood version of his view. In his reply, he stresses the distinction between *purely phenomenal spatiotemporal functionalism*, which is the view described in (3), and *combined phenomenal/non-phenomenal spatiotemporal functionalism*, which is the view he endorses. Purely phenomenal spatiotemporal functionalism claims that spatiotemporal concepts are characterized purely by their phenomenal roles. On the other hand, phenomenal/non-phenomenal spatiotemporal functionalism claims that spatiotemporal concepts are characterized by both their phenomenal and non-phenomenal roles. By adopting the latter view, D in Stoljar's argument is no longer a phenomenal dispositional truth, and D+ does not entail a phenomenal truth. Therefore, Stoljar's argument loses its force.

To strengthen his defence, Chalmers further makes the distinction between pre-theoretical physical concepts (p-physical concepts) and theoretical concepts (t-physical concepts).[41] P-physical concepts are characterized by their phenomenal roles, while t-physical concepts are only characterized by their non-phenomenal roles. According to Chalmers, for every p-physical concept (such as mass, space and time), we can create a twin t-physical concept (such as mass*, space* and time*). P-physical concepts and their twin t-physical concepts co-refer. This is where Chalmers's combined phenomenal/non-phenomenal spatiotemporal functionalism comes into play. Each physical property plays a partially phenomenal role and a partially non-phenomenal role. We can pick out the physical property via its phenomenal role using the relevant p-physical concept, but we can also pick out the physical property via its non-phenomenal role using the relevant t-physical concept. Furthermore, Chalmers argues that t-physical concepts are relevant in the discussion of metaphysics, while p-physical concepts are relevant in the discussion of epistemology. As a result, Stoljar's argument fails to bring out an inconsistency, as his argument can only run with p-physical concepts but not t-physical concepts. However, I do not totally agree with his distinction between p-physical concepts and t-physical concepts. First, I struggle to understand how certain physical concepts can be purely structural and lack any phenomenal features. The

[41] One should be careful not to mistake Chalmers's t-physical concepts here with Stoljar's t-physical concepts.

process of defining purely structural physical concepts appears unclear to me. Second, even if the distinction is drawn, Stoljar's argument can still run smoothly if p-physical concepts are used. We then need to conclude that p-physical truths entail certain phenomenal truths, which still seems to be an undesirable outcome for the dualists. With that being said, I will not discuss it any further, as it is not my goal to examine Chalmers's response in detail here.

3.2.3 My Reply

There are two replies I want to make about Stoljar's argument. The first one concerns the link between the strong no entailment thesis and the first premise of the zombie argument, and the second one concerns the link between D and D+. The first reply is similar to the one Chalmers has in his reply, but I will elaborate on it in my own way here. The second reply is completely unique from what Chalmers has mentioned. Chalmers defuses Stoljar's argument by endorsing combined phenomenal/non-phenomenal spatiotemporal functionalism instead of purely phenomenal spatiotemporal functionalism. However, I think even purely phenomenal spatiotemporal functionalism is perfectly compatible with the zombie argument, and I will argue why it is the case later. Let me start with the first reply here.

Put simply, my first reply is that the failure of the no entailment thesis does not entail the inconceivability of zombies. Stoljar of course foresaw this kind of objection. He makes a distinction between two statements (p. 477):

> (1) For any phenomenal truth T*, there is no physical or topic-neutral truth T, such that, T *a priori* entails T*.
>
> (1-w) For some phenomenal truth T*, there is no physical or topic-neutral truth T, such that, T *a priori* entails T*.

(1) is the extreme version of the no entailment thesis, and it is the one Stoljar used in his argument. (1-w) is a weaker thesis but it is still sufficient to threaten physicalism. Stoljar (2020) claims that Chalmers adopts (1) instead of (1-w):

> Chalmers himself endorses the former, and hence accepts the extreme version of the conceivability argument. The reason is that he accepts the conceivability, and so possibility, of zombie worlds, i.e., worlds that are identical to ours in respect of all physical truths, but at which there is no consciousness whatsoever. (p. 477)

This is incorrect. In Chalmers's reply to this argument, he stresses that (1) is not necessary for the zombie argument or dualism in general to succeed. Similar ideas can also be found in his earlier work (Chalmers, 2009). When (1-w) is adopted instead of (1), even if Stoljar's argument was sound, the zombie argument is not in such big trouble, as Stoljar's conclusion only shows

that some phenomenal truths are entailed by physical truths but not all phenomenal truths are entailed by physical truths.

Let me elaborate on this idea here. To adapt to (1-w), we can run a specific version of the zombie argument. Instead of conceiving of a complete zombie world where no phenomenal consciousness exists, we can conceive of a partial zombie world where whoever is running the argument (assuming that they are conscious) is the only conscious agent. If I am the subject that runs the argument, then I will conceive of a world where I am the only conscious agent. Conceiving such a world should be relatively easy, as we do not have direct access to other organisms' (not even other humans') phenomenal consciousness anyway, assuming that they have any at all. The possibility of such a partial zombie world still poses serious threats towards physicalism. Most physicalists, and perhaps most philosophers in general, believe that there are multiple conscious agents in the actual world. The logical possibility of a partial zombie world will at least ascertain the lack of *a priori* entailment from physical truths to phenomenal truths about others. The conceptual link between physical truths and phenomenal truths remains broken.

There is another reply Stoljar makes towards this kind of objection. Stoljar thinks that his argument does not merely demonstrate that one or some phenomenal truths are entailed by physical truths, but that a very large class of phenomenal truths. In particular, Stoljar claims that this class should at least include "any truth that reports a veridical spatiotemporal experience" (p. 477). However, from an epistemic standpoint, only phenomenal truths concerning the subject of reasoning can be included. The link between phenomenal truths and physical truths only exists due to the phenomenal elements in physical concepts. Therefore, it makes *prima facie* sense that my knowledge of the physical is *a priori* connected to my knowledge of my own phenomenal features. However, I do not have direct epistemic access to phenomenal information about others, even though I infer that phenomenal consciousness exists elsewhere in the world. Hence, even if Stoljar's argument was sound, the only phenomenal truths that are entailed by physical truths are phenomenal truths about myself.

But is Stoljar's argument even sound? I think not. My second reply attacks (P4) in the second phase of Stoljar's argument. As mentioned before, D is a phenomenal dispositional truth, and D+ is the conjunction of D and "x causes what it normally causes." (P4) then claims that D+ is *a priori* equivalent to a physical truth since D is *a priori* equivalent to physical truth. Stoljar finds (P4) compelling as D+ only contains further causal/nomic vocabulary compared to D. Furthermore, since the expression "is the normal cause of experience E" is built into D in

the first place, D+ entails a phenomenal truth – E occurs. Stoljar then concludes that there is at least a phenomenal truth that is entailed by a physical truth.

I do not agree that D+ is *a priori* equivalent to a physical truth, even if D is *a priori* equivalent to a physical truth. There is a question that brings up my doubts: If D+ is *a priori* to a physical truth, what physical truth is that? We have known from (C1) that the physical truth D is *a priori* equivalent to is S. We can then ask whether the physical truth D+ is *a priori* equivalent to is also S. There is a seeming contradiction if we answer yes. If both D and D+ are *a priori* equivalent to S, then we are forced to conclude that D is *a priori* equivalent to D+. But that cannot be the case, since D+ is the conjunction of D and "x causes what it normally causes." D and D+ are obviously two different propositions that are not *a priori* equivalent to one another.

We are then left with the option of saying D+ is *a priori* equivalent to a physical truth that is not S. Let us call this non-S physical truth S+. However, there are some undesirable consequences following this move. Suppose S is a physical truth about water. D should be a phenomenal dispositional truth containing "x is the normal cause of water-related experience." D+ is *a priori* equivalent to the conjunction of a phenomenal dispositional truth and a phenomenal truth, where the phenomenal dispositional truth is D and the phenomenal truth is "water-related experience occurs." The important difference here is that S (which is *a priori* equivalent to D) does not entail the phenomenal truth about the occurrence of water-related experiences, but S+ (which is *a priori* equivalent to D+) does. According to Stoljar, both S and S+ are physical truths, but what sort of physical truth is S+? We can first try to answer what truth S+ will be. The most natural answer is that S+ is *a priori* equivalent to the conjunction of S and "water-related experience occurs," which does not seem to be a physical truth, but a conjunction of a physical truth and a phenomenal truth instead. If one insists that both S and S+ are physical truths, one must consider there to be two very different kinds of physical truths, where one kind entails phenomenal truths and the other does not. This is counter-intuitive to say the least. My view here is that D and D+ cannot both be physical truths. More specifically, adding further causal/nomic expressions to D does not automatically make a new truth that is *a priori* equivalent to a physical truth. D-like truths and D+-like truths differ in a significant way. If we consider D to be *a priori* equivalent to a physical truth, then D+ cannot also be *a priori* equivalent to a physical truth. If D+ is not *a priori* equivalent to a physical truth, Stoljar's argument is undermined.

There is another worry concerning whether D+ will be a truth at all. Stoljar also thought of this concern and formulated a brief defence against it. He claims that the assumption of D+

being a truth "is a highly defensible assumption… if something is the normal cause of this experience, that thing plausibly caused what it normally causes" (p. 478). However, most non-panpsychist physicalists should agree on the existence of physical events that have never been experienced. There might exist a basketball-size rock two billion lightyears away that has never caused any experience. No one knows about its existence, but its existence nonetheless constitutes a physical truth. It is a physical truth because the rock possesses the disposition to generate specific experiences, even though such a disposition has never been realized. One way to avoid this problem is to make S as a known physical truth. After all, Chalmers's structuralism mainly concerns epistemology, not metaphysics. If that is the case, we can consider D+ instead of D to be *a priori* equivalent to S. What follows is that S entails a phenomenal truth that E occurs. However, E can only be an experience that the subject of reasoning has. We have direct epistemic access to our own experiences but not to others'. So, there is no entailment from physical truths to phenomenal truths about other systems' experiences. This brings us back to my first reply. If we adopt the weak no entailment thesis, the first premise of the zombie argument seems to be unharmed. So, it seems like we can either admit that D is *a priori* equivalent to a physical truth but D+ is not, or that D+ is *a priori* equivalent to a physical truth but only the strong entailment thesis is threatened. Either way, the first premise of the zombie argument is unthreatened. Combined with the first reply that I made, I conclude that we can endorse both structuralism and the zombie argument at the same time and that Stoljar's argument poses little threat towards the zombie argument.

3.3 Reply to Fischer and Sytsma (2021)

Most recently, Fischer and Sytsma (2021) launched an ingenious attack on the conceivability of zombies using experimental philosophy. They attempt to explain away the zombie intuition with the idea of linguistic salience bias and to demonstrate that laypeople do not even find zombies *prima facie* conceivable. More importantly, they conclude that the conceivability of zombies should be rejected based on their findings. In the following, I will first introduce the core ideas and the experimental design used by Fischer & Sytsma. Then, I will briefly go through their experimental results and the conclusions they draw from the results. At the end, I will argue that their findings fail to undermine the conceivability of zombies.

3.3.1 The Objection: Ideas and Experimental Design

Fischer and Sytsma's experiment is designed around one core idea: the linguistic salience bias. What is the linguistic salience bias? Fischer and Sytsma (2021) explain it by saying the following: "when encountering unbalanced irregular polysemes whose interpretation involves suppression, thinkers are liable to be swept along by stereotypical inferences, even when these are defeated by the context" (p. 4). So, in the case of the zombie intuition, the term "zombie" is usually associated with movie zombies by laypeople, where the stereotypical features of movie zombies include rotting bodies, aggressiveness, lack of consciousness, etc. At the same time, the philosophical usage of the term has a different sense or meaning, in which the two core features are being physically identical to humans and a lack of consciousness. When scenarios involving philosophical zombies are presented, people need to suppress all features associated with the common usage of the term "zombie," except for the feature of a lack of consciousness. Here, we can call the sense associated with the common usage of the term the dominant sense and the one associated with the philosophical usage of the subordinate sense. According to Fischer and Sytsma, due to the salience imbalances between the dominant and subordinate senses, bad inferences will arise in scenarios where the subordinate sense is required. The feature of the lack of consciousness associated with the common usage of the term "zombie" will facilitate the *prima facie* conceivability of philosophical zombies, even if they are not really conceivable.

I will then briefly introduce the core procedures and materials involved in their experiment. 638 participants were recruited online. First, they were asked to complete a set of demographic questions. They were then asked to read the following vignette:

> Imagine that in the future scientists are able to exactly scan a person's body, including their brain, at the molecular level. Using this information, they can then create an exact physical duplicate of that person's body and brain, molecule by molecule. The resulting ['zombie'/duplicate] will have a body and brain just like the original person's. The [zombie/duplicate] will also behave just like that person. But, when it comes to the [zombie/duplicate], all is dark inside.
>
> Imagine that scientists successfully scan and duplicate an average person in this way. What, if anything, do you think the resulting [zombie/duplicate] would be like? (Fischer & Sytsma, 2021, p. 5)

Roughly half of participants received the version of the vignette with the term "duplicate," and the other half received the version with the term "zombie." The vignette is intended to describe a scenario where philosophical zombies exist. The phrase "all is dark inside" is particularly

important here, as it suggests a lack of consciousness without using the term "consciousness." More about the decision to use the phrase will be discussed in detail later. Right below the vignette, nine items (questions) were presented on the same page:

> **A** The [zombie/duplicate] would be capable of having conscious experiences.
>
> **B** The [zombie/duplicate] would have an inner mental life, including feelings and emotions.
>
> **C** The [zombie/duplicate] would be sentient and experience its surroundings and sensations.
>
> **D** There is something it would feel like to be the [zombie/duplicate].
>
> **T1** The [zombie/duplicate] would have a rotting body and attack and eat humans.
>
> **T2** The [zombie/duplicate] would move slowly and have a lifeless face.
>
> **T3** The [zombie/duplicate] would lack free will and feel no joy.
>
> **A1** The [zombie/duplicate] would be capable of being sad and feeling hate.
>
> **A2** The [zombie/duplicate] would think and be intelligent.
>
> **A3** The [zombie/duplicate] would be capable of being happy, singing, smelling flowers, and feeling love. (p. 6)

Participants were asked to rate each item using a 7-point scale (1 = totally disagree to 7 = totally agree). Items A-D concern consciousness; T1-T3 concern the typical features of zombies; and A1-A3 concern the atypical features of zombies. All items were presented in random order. Furthermore, two comprehension checks were presented on the second page just to make sure that the participants understood that the zombies or duplicates are physically identical to their human counterparts. I will not present the materials here as the details are not important. Participants who failed either comprehension check were excluded from the data. 347 participants were left after the exclusion.

In this experimental design, one variable (term) is manipulated across two levels ("zombie" vs. "duplicate"). Two predictions were made by Fischer and Sytsma, with the first prediction being that participants in the "zombie" group will agree more strongly with T1-T3 but agree less strongly with A-D and A1-A3 compared to participants in the "duplicate" group. The second prediction is that more participants in the "zombie" group will disagree with A-D compared to the "duplicate" group. The two predictions are similar, with the major difference being that the first prediction focuses on the means of ratings while the second prediction focuses on the proportions of participants.

3.3.2 The Objection: Experimental Results and Interpretation

I will leave out the details of the results and data analysis, as they can be found in the original paper. Overall, the patterns in the results are clear-cut, with most statistical differences being significant. As predicted by Fischer and Sytsma, participants in the "zombie" group agreed more strongly with T1-T3 but agreed less strongly with A1-A3 and A-D compared to those in the "duplicate" group. Furthermore, more participants in the "zombie" group disagreed (gave ratings from 1 to 3) with A-D compared to the "duplicate group." Another finding from the experiment is the low proportion of participants that disagreed with A-D: roughly 30-40% of the participants disagreed with A-D in the "zombie" group, and merely 15-20% of the participants disagreed in the "duplicate" group.

In short, it shows that the choice of term ("zombie" vs. "duplicate") in the vignette had a big impact on how the participants answered the provided questions. More specifically, participants in the "zombie" group were more inclined to consider the individual presented in the vignette to possess typical zombie features such as having a rotting body and moving slowly, but less inclined to consider the individual to possess conscious experiences. Furthermore, most participants did not consider the individual to lack consciousness.

There are two main conclusions Fischer and Sytsma drew from these results. The first conclusion is that laypeople in general do not find philosophical zombies *prima facie* conceivable. The second conclusion is that linguistic salience bias can account for a significant proportion of the conceivability among those who do find philosophical zombies conceivable. In other words, some people only find the so-called "philosophical zombies" conceivable due to the feature of lacking consciousness associated with the common usage of the term "zombie." Last but not least, based on the two conclusions above, Fischer and Sytsma further claim that zombies are not really conceivable, as the apparent conceivability of zombies that some people have "rests on epistemically deficient intuitions" (p. 11).

3.3.3 My Reply

There is little doubt that Fischer and Sytsma's experiment reveals something about the influence word choice has on how we reason and answer questions. However, I disagree that the finding of the experiment has the power to undermine the conceivability of zombies. In my reply, I will focus on breaking the link between the results of the experiment and the implication Fischer and Sytsma drew from the results. I will first mention two minor problems concerning the experimental design. Then, I will point out the major issue concerning the phrase "all is dark inside."

The first minor problem concerns the use of positive conceivability in the experiment. Although positive conceivability is more powerful in bridging the gap between conceivability and possibility, the use of positive conceivability is not necessary for the zombie argument (Chalmers, 2002). Negative conceivability is often used by default when the zombie argument is being discussed (Chalmers, 2009). One advantage of using negative conceivability over positive conceivability is the simplicity of the concept: something is negatively conceivable if it is coherent, and something is positively conceivable if it can be coherently imagined, with imagination being a significantly more complex idea than coherence. Therefore, showing that zombies are *prima facie* positively inconceivable does not immediately undermine the argument. This, however, is a relatively minor issue.

The second minor problem concerns the sample used in the experiment. The majority of the participants are laypeople with little to no training in philosophy. A natural question to ask will be why laypeople's intuitions matter in philosophical questions instead of professional philosophers' intuitions. Fischer and Sytsma did conduct an analysis based on participants' demographic data and discovered that the few participants who found zombies conceivable were not in a better epistemic position. However, this analysis does not explain why significantly more professionally trained philosophers find zombies conceivable compared to laypeople. In Bourget and Chalmers's (2014) study, roughly 60% of respondents, who are all professional philosophers, considered zombies to be conceivable. Fischer and Sytsma also mentioned this study in their paper and seemed to be claiming that these intuitions from professional philosophers can also be explained away using linguistic salience bias. However, at the same time, conflicting findings were presented in their paper, with some supporting that philosophers are less prone to the bias while others supporting that philosophers can also be easily influenced by related framing effects. The evidence presented by themselves does not even support their conclusion convincingly. Overall, I think it would be a hasty decision to claim that philosophers' zombie intuitions can be explained away using linguistic salience bias, especially in the absence of a dedicated experiment. If that is the case, the results of the experiment fail to threaten the conceivability of zombies, as it is common sense that philosophers' intuitions should be prioritised in evaluating philosophical questions.

The major issue that is capable of undermining the conclusion of the experiment concerns the phrase "all is dark inside." Put simply, due to the way that the vignette and questions were phrased, the answers provided by the participants did not act as good indicators of the conceivability of zombies. The two core features of philosophical zombies are being physically identical to humans (=P) but lacking consciousness (=~Q). The vignette provided a

very detailed and explicit description of the condition P, using phrases such as “exact physical duplicate,” “molecule by molecule,” and “behave just like that person.” On the other hand, the condition of ~Q was only stated using a single metaphorical phrase – “all is dark inside.” At the end of the vignette, participants were asked what they thought the resulting zombie or duplicate would be like. Well, the simple answer to this question would be that the resulting zombie or duplicate is an exact physical duplicate of the person, but all is dark inside. But what does “all is dark inside” mean? This will then depend on the participants’ interpretations. The participants might interpret the phrase as expressing a lack of consciousness or as expressing something else. But as I will explain later, conceivability is irrelevant no matter how the participants interpret the phrase.

The real problem arises when we take into account the questions. Items A-D are all questions concerning whether the individual in the vignette has conscious experiences. Participants can respond with numbers from 1 (totally disagree) to 7 (totally agree). Fischer and Sytsma interpreted 1 as finding zombies inconceivable and 7 as finding zombies conceivable. However, I do not think this is the correct interpretation. When people are faced with these questions, the most natural response would be to find the answers in the vignette. If the participants interpreted the phrase “all is dark inside” as expressing a lack of consciousness, they would probably disagree with A-D. On the other hand, if the participants interpreted the phrase as expressing something else, they would probably be neutral or agree with A-D. The answers should not be interpreted as conceivability but simple comprehension of the vignette, especially of the phrase “all is dark inside.”

Why was such an ambiguous metaphorical phrase used in the vignette? Fischer and Sytsma (2021) provided a justification in their paper. According to them, the phrase is chosen here to negotiate the tension between the prevention of shallow processing and the accurate depiction of the scenario. Shallow processing might happen if the candidate phrase is overly explicit. For example, if phrases such as “lack of consciousness” or “has nothing to be like” were chosen, participants would most likely disagree with A-D as there are clear contradictions between the vignette and questions. At the same time, the candidate phrase must also accurately depict the scenario so that the participants understand the individual as lacking consciousness (~Q). The phrase “all is dark inside” was chosen as a compromise, but I do not consider it to be successful in doing its job. As mentioned earlier, if the participants had interpreted the phrase correctly, they would not have disagreed with A-D. The questions did not act as tests of conceivability, but instead acted like checks of understanding, which most participants failed.

Is there a way to modify the experimental design so that it tests conceivability? According to Fischer and Sytsma, the key to improving the design lies in finding a phrase that can successfully negotiate the above tension. However, according to my analysis, even if such a phrase can be found, the experiment still tells us very little about conceivability. Even though I am not against experimental philosophy, I cannot think of a design that can successfully examine conceivability using similar methods. Conceivability, especially positive conceivability, is a complex idea. Even if we directly ask whether the participants find a given scenario conceivable, the response might still heavily depend on how they interpret the term "conceivability." My suggestions for future research will be to focus on the idea of coherence first, as it is closely related to the idea of negative conceivability. We can present vignettes that depict incoherent scenarios to the participants and record their replies. Based on the data, we can attempt to discover the differences between people's replies when they are presented with a coherent scenario and when they are presented with an incoherent scenario. I have good faith in such a program being better at examining conceivability compared to Fischer and Sytsma's program. To conclude, even though Fischer and Sytsma's experiment provides some insights into the relationship between language and reasoning, it fails to pose any genuine threats to the conceivability of zombies.

Chapter 4: Locating the Roots of the Disagreements

As I was researching the debate on the zombie argument, there was one question always lurking at the back of my mind: why do we have so many disagreements surrounding this argument? Suppose all the premises in the argument are truth-apt. It must mean some philosophers are making mistakes about these premises while others are not. Philosophers are smart people, but why is it so difficult for them to tell what is correct and what is not? Why is it so difficult to reach an agreement? What are the roots of all these disagreements? Is it possible for philosophers to reach an agreement, especially on the soundness of the zombie argument, eventually? If we do reach an agreement, does that agreement resemble the truth?

This chapter will explore the above questions by discussing two topics: language and intuition. Of course, the above questions can be asked in most, if not all, philosophical discussions, but I think this does not discredit their relevance to the zombie argument. So, although they might not seem to be directly related to the argument, discussions on these questions can undoubtedly help us understand how we can truly settle the argument, and the mind-body problem in general. Overall, this chapter is better considered as a bonus part extended from the main body of the thesis and a place where I express my speculations on how potential progress can be made on the zombie argument.

Of course, the questions mentioned above are probably some of the most important yet difficult questions to answer in philosophy in general. I have no intention of solving them or providing a thorough analysis here. What I will try to do in this chapter instead is merely to provide my own reasoning and analysis of these questions. A lot of my ideas will inevitably overlap with pre-existing ideas, and I will attempt to mention these pre-existing ideas wherever I can. Despite the lack of a thorough analysis, I still consider it an interesting and meaningful way to end the thesis by exploring how these meta-philosophical ideas can be applied to the debate over the zombie argument.

4.1 Disagreements Based on Language

As discussed in Chapter 1.3, sometimes disagreements arise purely due to the confusion in the way we use our words, such as the case where Adam and Bob disagree on whether Carol likes orange, but in fact Carol likes eating oranges but dislikes the orange colour. This is a typical case of the verbal disputes that I will focus on in this chapter. In general, verbal disputes are disputes that can be fully resolved by simply specifying the meanings of terms or

expressions involved in the disputed statement.[42] Verbal disputes are interesting as the two (or more) parties in disagreement might not be disagreeing on the content of the statement at all but merely on whether a certain word or expression should be used.

In this section, I will first present a framework of language and concepts that I think best explains our philosophical intuitions and is capable of helping us resolve verbal disputes. The framework presented here will include, but not be limited to, ideas such as Jackson's (1998) conceptual analysis, the neo-descriptivism defended by Braddon-Mitchell (2004), and Chalmers's (2012) *a priori* scrutability. Again, I am not here to support or develop any specific view, but simply to present a view that makes the most sense. I will also briefly justify the use of the provided framework. Then, I will demonstrate how verbal disputes might be explained using this framework. Last but not least, I will discuss the potential verbal disputes in the discussion of the zombie argument and how they might be resolved or at least clarified.

4.1.1 The Ideal Framework of Language and Concepts

Let me start with the orange example again. Both Adam and Bob agree that Carol likes eating oranges but dislikes the orange colour. So, the dispute on "Carol likes orange" is purely verbal due to Adam and Bob attaching different meanings to the term "orange." In reality, we often attach multiple meanings to the same word. We call these words "homonyms." So, it seems to me that it is best not to consider words to carry meanings directly, but through some kind of medium. As mentioned in Chapter 1.3, I consider this medium to be concepts. Multiple concepts can be attached to the same word, but only one meaning can be attached to each concept. When I say meaning, I mean intension and extension, where the extension tells us about what the concept picks out in our world and the intension tells us about how the concept is picking them out. The use of intension allows two concepts to have different meanings even if they have the same extension. This is important for modal discussions, as two concepts might share the same extension in our world but nonetheless pick out different sets of things in another possible world. Furthermore, we can apply two-dimensional semantics to further distinguish between primary and secondary intensions. The primary intension of a concept picks out its extension in a world when it is considered as actual, and the secondary intension picks out its extension in a world when it is considered as counterfactual (for more information on two-dimensional semantics, see Chapter 1.3).

[42] A more detailed discussion on the definition of "verbal dispute" can be found in Chalmers 2011.

If concepts are the ones that carry meanings, the next question becomes how we can clarify the concepts involved in discussions. In other words, it is the question of how we can know about the meanings of words. Suppose someone says, "I like orange." How do we know what concept of *orange* they are using? It seems to me that the primary way to clarify the concepts is to rely on words again. We need to find a way to describe the concepts, and to do that, we need to understand the structure of concepts. Here, I adopt the classical theory where each concept has its own necessary and sufficient conditions that its referents must satisfy. The necessary and sufficient conditions can be described by a description, which in turn constitutes the definition of the term associated with the concept. I will briefly defend the use of necessary and sufficient conditions shortly.

Even if we have a way to describe the necessary and sufficient conditions, how do we know what the descriptions are? Here, we can apply Jackson's (1998) version of conceptual analysis. Suppose we want to find out what our concept of *bachelor* is (assuming that we all have the identical concept of *bachelor*). What we can do is to list a bunch of possible scenarios (or objects) and ask whether each (scenario) object intuitively falls into the extension of our concept. Once we have done that, we try to generate a description that best describes the pattern of division (between those that fall into and out of the extension). If such a description can be found, this description then successfully describes the necessary and sufficient conditions of the concept, and therefore becomes the definition of the term. For example, it is arguable that "unmarried man" is such a description for the concept of *bachelor*. At the same time, conceptual analysis can be used to reject a description. For example, conceptual analysis has shown that "justified true belief" is not the correct description for the concept of *knowledge*, as we can find possible scenarios where justified true beliefs are involved without the involvement of knowledge.[43] Now, one might notice that conceptual analysis seems to focus more on how words are connected to one another instead of concepts. This is partly true due to language being our most major means of communication. Even though we are trying to demonstrate the connections between concepts, we can only do so by showing connections between words that represent these concepts. Still, if we have a full understanding of the meanings of the descriptions, we should also have full understanding of the concept, as the descriptions reveal the necessary and sufficient conditions of the concepts.

Here comes another problem: what if I do not have a full understanding of all the terms in the description? Even if we have decided that being a bachelor simply means being an

[43] This example is most famously used in the Gettier problem. See Gettier, 1963.

unmarried man, we cannot take the meaning of "unmarried" and the meaning of "man" for granted. So, we will need to do the same conceptual analysis on the concept of *unmarried* and the concept of *man*. This leads to an apparent regress. Assuming that we have a finite number of concepts (and terms), there are two possible outcomes. The first option is to say that all terms can be defined in other terms, which means all the terms we have just define one another.[44] This seems to me to be a case that we have to avoid, as it leads to circularity.[45] The second option is to consider certain terms (and therefore concepts) to be at the bottom of the process. These terms cannot be further defined in other terms, which means the concepts they represent cannot be further broken down into smaller concepts. Chalmers (2011, 2012) calls them primitive or bedrock expressions (and concepts). These terms (or expressions) and concepts are primitive in the sense that they cannot be analysed in terms of other terms and concepts. Some candidates of primitive concepts might be phenomenal concepts (e.g., the concept of *phenomenal redness*), logical concepts (e.g., the concept of *truth*), and normative concepts (e.g., the concept of *intrinsic goodness*).

It is also important to understand that the meanings of words here are always *a priori* decided by the users themselves. This means the users have complete authority over what concepts are represented by what words, and what the meanings of the concepts are. I can freely choose to use the word "apple" to mean *unmarried man*, as there is no objectively correct meaning for each word. However, this is obviously not pragmatically desirable. So, most of the time, we seek to align our meanings of words and have the same concept represented by the same word. One way to do it is to utilise deferential concepts. For example, I understand that most people take biology as the authority to tell us what is an apple and what is not. So, to make sure my concept of *apple* picks out the same extension as most other people's concept of apple, I might use "apple" to mean *whatever the biologists use the word "apple" for*. Now, it might seem like the meaning of "apple" has become an *a posteriori* matter, but this is not entirely true. The true intension of my concept of *apple* is *whatever the biologists use the word "apple"*, which is still entirely *a priori* decided by myself. Still, suppose both my friend and I

[44] Of course, every term necessarily defines itself in a trivial way. Here, I mean the non-trivial kind of defining, where a term must be defined using other terms.

[45] Sometimes certain apparent circularities might not be vicious. According to the method of Ramsification (mentioned in Chapter 3.2, also see Lewis, 1970), theoretical terms within a theory might be defined in terms of one another. For example, t_1 might be defined as "the thing that orbits around t_2," and t_2 might be defined as "the thing that t_1 orbits around." However, the circularity here is non-vicious. In fact, we can define the pair [t_1, t_2] as the pair that satisfies such-and-such conditions, where such-and-such conditions are explicated without using the terms "t_1" and "t_2." Thanks to David Braddon-Mitchell for the discussion here.

have the same concept of *apple*, which is *whatever the biologists use the word "apple" for*; we can still disagree on the way biologists use the word "apple" for. In other words, we might have a dispute on "X is an apple" purely due to our disagreement on what biologists mean by "apple." The two parties in this case agree upon the meaning at one level but disagree at another level. Chalmers (2011) calls disputes of this kind "broadly verbal disputes." I prefer considering disputes of this kind not to be purely verbal, as the two parties in fact disagree on the *a posteriori* empirical fact of what the biologists mean by the word "apple." For now, I will focus on narrowly verbal disputes, which are those where two parties completely disagree over the meaning of the word.

Just to quickly review, so far, we have a framework where words (or terms or expressions) represent concepts, and concepts carry meanings. To have a meaning is to have an intension that is capable of picking out an extension. At the core of the intension is a set of necessary and sufficient conditions that all things in the extension must satisfy. To clarify what the necessary and sufficient conditions are, we have to use a description, which in turn constitutes the definition of the term that represents the concept. Furthermore, the necessary and sufficient conditions of the concepts, so as the definitions of the words, are *a priori* accessible to the users of the words and concepts.[46]

There have been plenty of objections to the classical theory of concepts, along with the use of definitions, descriptions, and necessary and sufficient conditions in general. According to Laurence and Margolis (1999), some of the main criticisms of the classical theory include Plato's problem, the problem of psychological reality, the problem of ignorance and error, the problem of conceptual fuzziness, and the problem of typicality effects. However, none of these criticisms seem to provide a knockdown argument against the classical theory. Many defences have also been made against these criticisms (e.g., Jackson, 1998; Chalmers & Jackson, 2001; Braddon-Mitchell, 2004; Chalmers, 2012). Of course, properly defending the classical theory can be an arduous and complicated process, which I have no intention of getting into here. Instead, I will briefly focus on one kind of objection – objections that rely on arguing that the classical theory fails to explain how concepts and words are actually used by people. These objections often point out certain empirical data that cannot be explained by simply appealing to necessary and sufficient conditions. For example, according to the problem of typicality effect, people usually consider certain instances of a concept to be more typical than others.

[46] If that is the case, why do we still need dictionaries? Dictionaries are still useful when we want to know about how a word is used by the linguistic authority in our society. This kind of information is specifically useful if we have deferential concepts for our words.

Another example might be the lack of explicit analysis, where accurate explicit definitions rarely ever come up for any concept. If people's concepts are characterised only by necessary and sufficient conditions, it seems like these empirical phenomena should not have occurred. So, these empirical findings act as evidence against the classical theory of concepts. The question now becomes why we should adopt the classical theory if there are alternative theories, such as the prototype theory, that do a better job at explaining how concepts and words actually work.

How should we reply? I think we should simply concede that the classical theory (along with the use of definitions, descriptions, and necessary and sufficient conditions) does not do well in explaining how people actually communicate and think. Instead, the classical theory describes how concepts would be used if people were perfectly rational, which we are not. To be evolutionarily advantageous, humans simply need to be good enough at interacting with the environment with the goals of survival and reproduction. So, the ways we think and communicate are likely to be shaped to serve evolutionary purposes and not philosophical purposes. The question of how humans conceptualise things in the real world remains a mysterious one, and answering this question is the job of psychology, not philosophy. So, I will instead provide some justifications on why I think the framework provided above, which is based on the classical theory of concepts, works best with ideal rational reasoning.

The key assumption I am using here is that concepts should be capable of distinguishing between things (or properties) that fall into or out of their extensions. This feature is very important for statements to be truth-apt. For the statement "all apples are red" to be truth-apt, for example, we need the concept of *apple* to be able to distinguish between apples and non-apples in our world. For modal statements such as "apples are necessarily red" to be truth-apt, we need the concept to do the same thing in other possible worlds as well. Otherwise, any non-red thing that neither falls into nor out of the extension of *apple* will be a problematic case for our statements. Furthermore, concepts do not just make the distinctions randomly; things that are in the extension must share some common features that distinguish them from things out of the extension. It seems to me that the only way to do this is to rely on necessary and sufficient conditions: something falls into the extension of our concept if and only if it satisfies such-and-such conditions. The necessary and sufficient conditions here exist as the natural consequence of the clear-cut extensions of concepts. Then, to clarify what the conditions are in words, we need to come up with a description for the "such-and-such conditions." The description here then acts as the definition of the term, assuming the term only represents one specific concept.

In conclusion, I think for sentences and propositions to be truth-apt, the above framework is the most natural answer. It is the perfect tool for philosophical investigations. Humans are not perfectly rational thinkers, but in philosophy, we always aim to get as close as possible to ideal reasoning. So, it makes sense to have a distinction between a theory of concepts that explains how people actually conceptualise the world and a theory of concepts that tells us how we should conceptualise the world if we want to achieve near-ideal reasoning. The framework here counts as a candidate for the latter. I do realise I have only provided a very brief defence of the framework here, and readers might still disagree that this is the framework that we should be using. If that is the case, the framework itself can simply be considered as an assumption that I adopt here for the purpose of my discussion.

4.1.2 Explaining Verbal Disputes Using the Provided Framework

In the previous section, I have already teased the relationship between our framework and verbal disputes. In this section, I will provide more details on how such a framework can explain and clarify verbal disputes. The first question, however, concerns whether verbal disputes can even be explained by the provided framework. As mentioned earlier, this framework intends to explain how we should think and talk, not how we actually think and talk. If so, what is the point of analysing actual disputes using the framework? There are two replies I want to make. First, even though it is unlikely that people actually possess clear-cut necessary and sufficient conditions for their concepts, we can often come up with approximate necessary and sufficient conditions associated with the ways they commonly use their concepts. Sometimes, disputes arise due to these approximate necessary and sufficient conditions being significantly different, and in these cases, the provided framework can be useful in providing us insights into the nature of the dispute.

Second, I think it is of philosophers' interest for concepts to have necessary and sufficient conditions. Here is the reason. A dispute over a statement occurs when one party thinks the statement is true and the other thinks it is false. For that to happen, the statement must be truth-apt, and as discussed earlier, necessary and sufficient conditions must be involved. Of course, if we adopt this idea, there is the important question of how to make sense of statements in everyday life: if most concepts possessed by regular people do not contain a set of clear-cut necessary and sufficient conditions, are statements containing these concepts truth-apt? This is certainly an interesting question that awaits to be answered, but I will not get into the details due to the scope of this thesis. Personally, I think interpreting statements containing concepts that lack necessary and sufficient conditions will be an extremely difficult,

if not impossible, task. So, I think philosophers should at least aim to come up with necessary and sufficient conditions for the core concepts involved in their discussions, and we should first understand how to analyse disputes over statements that contain concepts with necessary and sufficient conditions. Given the two reasons above, I will only be dealing with disputes that involve concepts that are compatible with our framework.

After making the above assumptions, verbal disputes, especially narrowly verbal disputes, occur when people associate different concepts with the same word.[47] We can also say that they use the same word to mean different things. To solve the problem, we simply need to clarify the meanings of all the words used in our disputes and identify the differences in how the words are used. Suppose two parties have a dispute on a statement S, and X is the word of interest. According to Chalmers's (2011) method of elimination, we can ban the usage of X, and let both parties clarify what they mean by X. We might discover that one party uses the term X to mean X_1 and the other uses X to mean X_2, where X_1 and X_2 are the two definitions of X used by the two parties accordingly.

But how do we come up with X_1 and X_2? If the two parties are both explicitly aware of the ways they are using X, they should have clear access to X_1 and X_2 accordingly. But very rarely in real life do we have explicit knowledge of how we are using our words. I might have intuitive judgements of whether X is involved in any given scenario, without knowing what sort of descriptions or criteria best explains my intuitive judgements. If this is the case, we will need to rely on Jackson's (1998) conceptual analysis. If the two parties use X differently, the conceptual analysis will reveal certain scenarios where one party thinks X is involved and the other party thinks otherwise. In an ideal situation, with a thorough conceptual analysis, we will find two descriptions that best describe how the two parties make their intuitive judgements related to X. The two descriptions here will be X_1 and X_2 accordingly. Once we have done that, we can then rephrase S into two separate statements, S_1 and S_2, using X_1 and X_2 correspondingly. As long as both parties agree on the truth values of S_1 and S_2, we can say that the dispute on S is wholly verbal.

[47] Chalmers focuses on broadly verbal disputes instead in his 2011 paper. Broadly verbal disputes often occur with deferential concepts, where the two parties disagree on the way their linguistic community uses a word. Here, it seems to me that it assumes the existence of an objectively correct community meaning of the disputed word and that the two parties simply disagree on the empirical fact of what this community meaning is. Therefore, I do not consider the two parties to genuinely disagree on the meaning of the word, as they both use the word to represent the same deferential concept. This is why I focus on narrowly verbal disputes here.

However, I think this is not the end of it. First, how do we know if the two parties agree upon the truth values of S_1 and S_2? Take S_1 as an example. We know that both parties truly agree upon S_1 only if they both understand S_1 in the same way. Otherwise, it might be the case where one party interprets S_1 as S_{1a} and the other interprets S_1 as S_{1b}. The agreement here is superficial, as there might be substantial disagreements that are hidden from the verbal misunderstandings. To avoid this, we need to clarify all the terms involved in S_1, which means introducing more terms in the process of clarifying the terms in S_1. This process will continue until we have only primitive terms left or until we simply assume the two parties agree upon the meanings of all the terms involved at any given stage of the process.

What happens when we reach the bottom? There is no further clarification available. Sometimes, we can figure out whether people possess the same primitive concepts by observing how people apply their concepts in different scenarios. In these cases, we might still gain an understanding of what concepts other people possess, even though we have no description to describe them. However, if certain primitive concepts have no behavioural influence, such as phenomenal concepts, we will have no way to confirm the content of these concepts. We simply need to assume that somehow we attach the same primitive concepts to the same words. Most of the time, I will say that our assumptions are at least reasonable. For example, when I ask my fellow humans what the colour of ripe tomatoes is, they answer "red." I expect most human beings to share the same visual experience as I do when looking at a ripe tomato, which means they share the same concept of *red experiences*. So, I make a reasonable assumption that other English-speaking humans use the word "red" to mean the same thing as I do with the word "red." Here, I no longer rely on descriptions to clarify concepts because primitive concepts lack definitions. Instead, I rely on the assumption that people evoke the same concepts in certain situations. If we accept the bigger assumption that people share all the same primitive concepts and that we also have a way to communicate these concepts, we will be in principle able to clarify all terms and eliminate all verbal disputes using conceptual analysis. However, this assumption certainly should not be taken for granted.

In reality, we rarely go all the way down to the primitive concepts and terms. One of the main reasons is that although the concepts we express with each word are rarely identical, they are often close enough to pick out a similar extension. They are good enough for everyday purposes. However, in philosophical discussions, we should be extra careful. The exact meanings of terms and sentences are of great interest in philosophy, and they are crucial to the development of philosophical discussions. Furthermore, misunderstandings in philosophy happen very often due to the unique and specific ways philosophers use words. Due to the

special nature of philosophical investigations, words are often used very differently in philosophy compared to everyday life. At the same time, these misunderstandings can significantly impact, in a bad way, how philosophy progresses.

There are two main implications from the above analysis. First, we can rarely be certain about the meanings of sentences from others, if they have any meanings at all. Even if we keep asking the other person to clarify their words by providing definitions, we will eventually run into primitive terms that express primitive concepts that cannot be further clarified. This problem becomes more severe with concepts that have no behavioural influence, such as phenomenal concepts. I can never be certain about whether other people evoke the same concept of *phenomenal redness* when looking at ripe tomatoes, as I have no direct access to others' phenomenal states. Second, even though we cannot ascertain the meanings of others' sentences, we can often get pretty close. With the help of conceptual analysis and the method of elimination, we can still gain decent insights into the meanings of the propositions other people are trying to express. So, when a verbal dispute occurs, we can often spot it. And if we assume that people share the same primitive concepts and that we somehow have a way to communicate these concepts, then we can in principle help each other understand what we are trying to express. In any way, the analysis above seems to provide a very natural and intuitive way of understanding verbal disputes.

4.1.3 Verbal Disputes and the Zombie Argument

It is not difficult to see how verbal disputes can arise in the zombie argument. Many terms involved in the argument can be used to mean different things, and agreeing upon the meanings of these terms is crucial for settling debates concerning the argument. For example, whether zombies are conceivable heavily depends on what "conceivable" means. Verbal disputes can easily arise if the term "conceivable" is not properly clarified. The same goes for some other terms or phrases in the argument. In this section, I will focus on discussing potential verbal disputes concerning the term "physical" and the term "conscious." After that, I will share some of my own speculations on how we can make better progress on the argument in general by clarifying our language and concepts.

The zombie argument is commonly considered an argument against physicalism. The success of the argument naturally relies on the kind of physicalism it attacks. As mentioned in Chapter 1.6, the kind of physicalism that we get mainly depends on the formulation of the thesis and the concept of *physical*. Here, I will focus on the concept of *physical*. We know that the kind of physical concept used in generating physicalism matters a lot in the zombie

argument. The zombie argument, especially the two-dimensional version, only denies t-physicalism but not o-physicalism, where t-physicalism adopts the t-physical conception and o-physicalism adopts the o-physical conception. Zombies are definitely secondarily possible if t-physical concepts are used, but might be secondarily impossible if o-physical concepts are used. Two parties can have a verbal dispute on the secondary possibility of zombies without having any real substantial disputes. This is a case where potential verbal disputes can be resolved by simply clarifying the concept of *physical* in use.

However, the way the physical concepts are clarified here can seem a bit forced. According to Stoljar (2001a, 2022), t-physical properties are the properties that our physical theory picks out, and o-physical properties are the properties possessed by the typical physical objects. Using the framework provided above, we can further ask what the "physical theory" and the "typical physical objects" are. Although we can avoid circularity by saying something like "the physical theory is the theory from the discipline called 'physics'," I find this explanation unsatisfactory, or inelegant at the very least. This sort of definition fails to explain what the physical theory is really about. So, although making the concept of *physical* non-primitive allows us to capture the potential different understandings of the term "physical," finding a satisfactory definition for each usage of the term is certainly no easy task. The lack of a satisfactory definition here can be a major source of confusion. So, progress can be made on the argument simply by coming up with a satisfactory analysis of the concept of *physical*.

If we suppose the t-physical conception and the o-physical conception are satisfactory, many of the verbal disputes concerning physicalism can be resolved by simply pointing out this distinction. We can simply say that t-physicalism is definitely false, but o-physicalism might still be true. But can we say that physicalism in general might still be true? If we consider o-physicalism as a kind of physicalism, the answer will be yes, but I think we need to be more careful here. There is a certain freedom in how we like to name a specific view. After all, names are often only considered as mere labels. But at the same time, names can have significant influence on how we behave and think about things. Most people associate the term "physical" with t-physical concepts. So, when people see the word "physicalism," many people might instinctively have t-physicalism in their minds. This effect might be especially significant for people who are not familiar with the t-physical/o-physical distinction. Combined with the reasons given in Chapter 2.3, I think it is best to avoid saying that the secondary possibility of zombies leads to physicalism for the purpose of preventing misunderstandings. Also, it can be argued that the term "Russellian monism" better describes the view compared to "o-

physicalism." Although it is not technically wrong to say that o-physicalism might be true, there are practical reasons why we might want to avoid the term "physicalism" in our case.

I will now turn to the term "consciousness." The term "consciousness" faces a different issue compared to the term "physical." We have trouble further clarifying "consciousness" due to the primitive nature of the term and its associated concept. There would not be any issue if people just agreed upon what "consciousness" means, but this does not seem to be the case. Most notably, analytic functionalism considers "consciousness" to be defined in functional terms, but the zombists clearly are not interested in the functionally defined concept of *consciousness*. If we adopt this concept of *consciousness*, zombies are plausibly inconceivable, even to the zombists. So, to reply, the zombists will point out that it is the concept of *phenomenal consciousness* that they care about, not the concept of *psychological consciousness* which is functionally defined. Does that clear the potential confusion? Not completely. Analytic functionalists consider psychological consciousness to be the only kind of consciousness that exists. So, when the zombists talk about the non-functionally defined "phenomenal consciousness," analytic functionalists will deny that the term is even capable of referring to anything real. At the same time, zombists, including myself, think that the existence of phenomenal consciousness as one of the most indubitable truths.

Is it possible that the analytic functionalists and the zombists are simply disagreeing on the meaning of "phenomenal consciousness?" Maybe. If this is the case, the problem can then be resolved once the zombists clarify what they mean by "phenomenal consciousness". However, we immediately run into a problem as the concept of *phenomenal consciousness* is arguably primitive, which means it cannot be clarified using any description or definition. We can still try to clarify it in other ways. If someone asks me what I mean by the term "phenomenal redness," I might show them a piece of red paper and say, "It is what you visually experience when you look at this." Still, here I am assuming that they have the same understanding of "visual experience" as I do and that they do indeed share the same experience when looking at that piece of paper as I do. The problem is not properly resolved. This is why primitive terms and concepts are so tricky to deal with. Very often, we simply need to assume that people possess the same primitive concepts and that they use the same words to express these concepts. In the case of analytic functionalism, this assumption cannot be taken for granted. From my perspective, if a conscious agent truly understands what I mean by "phenomenal consciousness," they surely will agree that it refers to something real. But this is not what the analytic functionalists say. This is probably why it is so difficult to argue against

analytic functionalism, despite its issues being obvious in the eyes of the zombists. For now, I have no answer to it myself.

A similar story can be told about some representational theories of consciousness, in which consciousness is defined as specific representational states. If these theories are adopted, zombies might not even be conceivable. In this case, we might be able to apply the subscript gambit and say that consciousness$_1$ logically supervenes on the physical but consciousness$_2$ does not, where consciousness$_1$ is the concept representationalists use and consciousness$_2$ is the concept zombists use (for subscript gambit, see Chalmers, 2011). Of course, just like what I have written above, the representationalists can still deny that consciousness$_2$ is capable of referring to anything real.

There are certainly some other disputes over the zombie argument that are potentially verbal. The disputes involving the terms "consciousness" and "physical" are only the two cases I want to focus on here. Sometimes it can be difficult to identify whether a dispute originates from misunderstandings of words, but we should always be on high alert. When we analyse any disputes, we should always keep in mind the possibility of the dispute being caused by two parties using the same word in a different way.

The last thing I want to talk about here is how the framework mentioned earlier can be used to make general progress on the argument. First, we should clarify any ambiguous terms (and concepts) used in the argument. The best-case scenario is to come up with clear-cut definitions of these terms. This will allow the audience to understand the exact meaning of the term with no remaining ambiguities, assuming that all the terms used in the definition need no further clarification. Definitions will also clarify the *a priori* relationship between the relevant concepts. For example, if conceivability is defined in terms of coherence, we can easily explain why conceivability is a good guide to logical possibility, as both concepts *a priori* rely on the concept of *coherence*. However, even if we successfully come up with seemingly satisfactory definitions, we will still need to further clarify the terms used in the definition. Ultimately, we will be left with primitive terms that cannot be further clarified using definitions. This will no doubt be an issue for us, as we cannot use the same method to make sure that we understand the primitive terms in the same way. Nonetheless, the attempts to provide close descriptions and definitions in most cases still help us form a better understanding of what other people might mean by their words and eliminate many of the potential verbal disputes involved.

Due to the problems with defining primitive terms and coming up with explicit definitions, we can also choose to clarify concepts indirectly, such as by providing scenarios where the concept represented by the term is involved and scenarios where the concept is not

involved. Although these indirect ways might not reveal the most exact meaning of the term as they do not clearly specify the necessary and sufficient conditions of the concept involved, they can often work quite well in delivering an approximate meaning. In any way, clarifying the concepts is important, and much work has already been done. For example, Chalmers himself has provided an extended clarification on the concept of *conceivability* in his 2002 paper, and clarifications on many other important concepts elsewhere (see Chalmers, 1996, 2006a, 2006b, 2009). It is then important for the readers who are interested in the argument to realise that these works exist and that these works are key to how the argument should be understood. Many apparent disputes surrounding the argument will resolve themselves once the argument is understood properly.

4.2 Disagreements Based on Intuitions

Not all disputes are verbal disputes. Two parties can disagree on the truth value of a sentence even when both parties understand the sentence in the exact same way. We can call these disputes "substantial disputes." Most of the time, philosophers seem to care more about substantial disputes than verbal disputes. When one philosopher critiques the view of another philosopher, the critic usually assumes that they understand the exact content or meaning of the view, but disagrees with it nonetheless. Understanding how substantial disputes might arise will no doubt help us understand how progress can be made more reliably in philosophy.

In this section, I will share my thoughts on the relationship between substantial disputes and intuitions. More specifically, I will talk about how substantial disputes might arise due to differences in intuitions. First, I will clarify the specific notion of intuition in play here and explain how intuitions of this sort are related to substantial disputes. Second, I will discuss the implications our analysis has on philosophical progress. Lastly, I will discuss how progress can be made on the zombie argument using our analysis of intuition.

4.2.1 What is the Intuition of Interest Here?

Commonly, intuitions can be roughly understood as things that seem instinctively true without requiring conscious rational reasoning. For the purpose of our discussion, I want to deal with a more specific kind of intuitions here, which will be defined as the following: *intuitions are beliefs about the truths of some propositions that cannot or do not need to be further justified*. As intuitions here are understood as beliefs, they are subject-related – what appears intuitively true to me might not appear intuitively true to others. Chalmers (2014) points out that when we find something intuitively true, we tend to have some kind of intuitive

justification for our intuitions – a sense of obviousness about the truth of the proposition that we believe in. For Chalmers, this kind of intuitive justification is significantly different from the common justifications that we have, which he calls "broadly inferential justifications." If we take Chalmers's view into account, we can slightly tweak our definition by saying that *intuitions are beliefs about the truths of some propositions that cannot or do not need to be further justified by any broadly inferential justifications.* The difference between the two definitions does not matter much for our discussion. In the following, unless I specify otherwise, when I use the word "intuition," I mean the intuition defined here. Although the definition here is not trying to capture exactly what people mean by "intuition" in philosophy, it should at least capture a very similar concept. When people say they find a proposition intuitively true, it is usually the case that the proposition appears true to them without the need for any (broadly inferential) justifications.

To demonstrate why intuitions matter so much in philosophy, let me first explain why unjustifiable claims very likely exist. Suppose we want to justify a claim A. To justify A, we must provide evidence or reasons in favour of A. In the process of providing the evidence, we will need to introduce at least one other claim B. Usually, people will say something along the lines of "since B is true, A is true" or "because of B, we should believe in A." Let C be the claim that "B provides good support for A." To successfully justify A, we want both B and C to be true. But why should we believe in B and C? Either we assume that B and C are true, or that we need to justify B and C. To justify B and C, we simply repeat the process above. In the process of justifying B and C, we will need to introduce more claims, such as D, E, and F, which in turn demand justifications for themselves. This process can be repeated over and over again. Suppose we only have a finite number of claims or propositions.[48] We are left with only two options: either we fall into circularity, letting all our claims justify one another, or we simply assume the truths of certain claims and no longer demand justifications for them. I suggest that the circularity option should be avoided for obvious reasons. Of course, proponents of coherentism might consider this option to not be as repugnant as it sounds. However, I think the problems with coherentism outweigh its benefits, with the biggest problem being the possibility of having a set of false or even ridiculous beliefs justifying one another. Although I

[48] It might be argued that we can potentially generate an infinite number of propositions by combining smaller propositions using logical operators. However, it seems to me that propositions generated in this way are unlikely to play an influential role in our belief system. Our beliefs in these complex propositions should be more or less determined by our beliefs in the more basic propositions.

am not planning to thoroughly argue against coherentism, for the sake of our discussion, I will rule out this alternative here. Once we rule out circularity, we end up with claims that do not demand justifications but nonetheless are capable of acting as justifications for other claims. When we are asked why we think these unjustified claims are true, the only answer we have is that we find them intuitively true. Therefore, unjustifiable claims exist, and their intuitive appeals act as the only reason we consider them true. Furthermore, since sometimes people share different intuitions, whether a claim appears unjustifiable and whether an unjustifiable claim appears true or false will also vary from individual to individual.

We should all be familiar with the above process, as that is basically how arguments work. Suppose we have a valid argument with a few premises and a conclusion. The opponents of the argument can argue against the argument by questioning one or a few of the premises. The proponents of the argument then need to point out the issues in the objections and justify how the questioned premise is true. This involves introducing other claims or propositions. For example, I (as a proponent of our argument) might build a new argument arguing for the truth of the questioned premise or a new argument arguing against the objection. But then, the premises of these new arguments can also be called into question. The process repeats until we reach a stage where we stop questioning our claims or propositions and their truths are taken for granted. This seems to be how mathematics works. All theorems and calculations are based on a restricted finite set of axioms that most mathematicians agree upon, despite the lack of proofs and justifications.[49] We consider the axioms as true simply because they intuitively appear so, and these axioms are used to justify and support everything else.[50]

Of course, not all intuitions are unjustifiable. Some claims can be easily justified but count as intuitions nonetheless due to their strong intuitive appeals (i.e., they appear obviously true without any broadly inferential justifications). An example of such claims might be "philosophers are smart." Some other claims might seem unjustifiable only because we have

[49] I am not saying that no mathematician doubts these axioms. Of course, debates exist over some of these axioms, but I do not believe it is the case that most axioms are debated among the majority of the mathematical community. There should still be a set of axioms that most mathematicians agree upon.

[50] One might argue that some widely accepted axioms might not seem intuitive at all. However, the question now becomes why we accept these axioms as true. The most natural answer I can think of is that there are some more influential intuitions that compel us to believe in them. Clashes of intuitions happen sometimes, and something might appear *prima facie* counter-intuitive but become intuitive after some thinking. Thanks to Michael Nielsen for pointing out this concern.

not thought about them hard enough. An example of such a claim might be "1+1=2."[51] So, certain claims count as intuitions despite being justifiable. However, I think scientists and philosophers should always at least attempt to justify their beliefs using broadly inferential justifications whenever possible, no matter how obvious the beliefs are. The claims that interest me the most are the ones that we struggle to find justifications for. We have to decide what to do with them. Many of these claims not only appear obviously true to us but are absolutely essential in helping us justify our other beliefs. The existence of these unjustifiable or difficult-to-justify intuitions poses interesting questions.

What are some examples of unjustifiable intuitions? One potential example I can think of involves logical rules such as "If either A or B is true, A and B cannot be both false." Such logical rules appear obvious to most people, and as far as I know, there is no proof or justification for why they are true. Of course, I cannot guarantee that they are truly unjustifiable. However, the existence of unjustifiable claims is not just supported by examples, but mainly by the analysis before. If we somehow find a way to justify these logical rules, we will then need to justify the claims that are used to justify the logical rules. We will eventually end up with unjustifiable claims.

What is the relationship between intuitions and disputes? Suppose two parties have a substantial dispute, as opposed to a verbal dispute, on a claim G. For any progress to be made on the dispute, both parties will need to justify their positions. In other words, one party needs to provide an argument for G, while the other needs to provide an argument against G. If one of the two parties provides a convincing argument that both parties agree upon, the dispute is resolved. For this to happen, the two parties must agree upon all the premises and the validity of one of the two arguments. This is relatively rare in philosophy, especially considering the fact that many philosophers are defensive about their own views. What often happens is that the proponents of G will point out one or multiple issues in the opponents' argument, usually by arguing against one or multiple of the premises. Suppose the premise being questioned is H. The two parties now have a dispute over H. The process then repeats. The process stops either when we stop asking for justifications and just accept whatever claims we end up with at a certain stage or when we reach claims that are unjustifiable. If the two parties disagree on an unjustifiable claim, the dispute becomes unresolvable. One party simply finds the claim intuitively true, while the other finds it intuitively false, with no justification or argument

[51] "1+1=2" can be proved using Peano's theorem. Thanks to Peter Godfrey-Smith for pointing this out.

available to settle the issue. On the other hand, if the two parties agree upon the unjustifiable claims, we can take the claims for granted and build our agreements utilising these claims. Again, there is no argument or justification for why they are true. We simply assume that they are true, as most people find them intuitively true. In conclusion, substantial agreements and disagreements very often come down to how people's intuitions, justifiable or unjustifiable, align and differ.

4.2.2 Intuitions and Philosophical Progress

We have seen earlier that there are plenty of disputes concerning the zombie argument. As philosophers, especially those who work on the argument, we want to resolve these disputes and make progress on the argument. But to do that, we first need to understand how philosophical progress can be made in general. In this section, I will focus on making some rudimentary analyses of how intuitions might influence philosophical progress. Before I start the analysis, we first need to define progress in philosophy. I think Chalmers (2015b) gives an intuitively appealing definition of philosophical progress, which is "large collective convergence to the truth" (p. 4). Although some people might not consider getting to the truth as the sole goal of philosophy, it is arguably the most important goal, and it is not enough to only let a minority of academics get to the truth. For many big questions in philosophy, there are usually multiple possible answers, each supported by their proponents, and it is likely that one of the answers is correct or true. This does not mean much to the discipline of philosophy in general. So, to say that philosophy in general is making progress, we must have most of the workers or academics reach an agreement that closely resembles the truth. After defining progress in philosophy in this way, it is not difficult to see why understanding disputes matters so much, as the existence of disputes suggests the failure of collective convergence to the truth.

I have talked about the relationship between intuitions and substantial disputes earlier, and I think this in part explains why disagreements are so ubiquitous in philosophy. In mathematics, almost all theorems are based on a set of widely accepted axioms that appear intuitively true to most people. So, agreements over these axioms will lead to widespread agreements across mathematics. In philosophy, there is no such set of axioms. Questions in philosophy are often based on different kinds of intuitions that are not widely shared among philosophers. Chalmers (2014) points out that "the use of intuitions in philosophy is more extensive, more focal, and more subject to disagreement than in many other fields" (p. 543). The widespread disagreements across philosophy can be explained by the unique ways of intuitions being used in philosophy. I do not think there is yet a satisfactory explanation of why

intuitions in philosophy are so contentious, and finding such an explanation will no doubt encourage progress in philosophy.

Let us go back to the relationship between intuitions and philosophical progress. Progress is made when disputes are resolved and the final agreement of the dispute reflects the truth. I will focus on the former part first. As mentioned earlier, there are two resolutions substantial disputes might end up with. First, if an agreement has been reached at some stage of the justification process, the dispute will be resolved. Once enough explanations and clarifications are provided, the two parties realise they share the same intuitions. One party will usually admit that they have made a mistake in their reasoning somewhere. Second, there might be disputes over the fundamental unjustifiable intuitions. Disputes of this sort cannot be resolved by providing justification or explanation for either position, as there is no more available justification to use. It is arguable that progress is made in the first case as long as the agreed-upon position reflects the truth. The second case is a bit trickier, as the dispute is not properly resolved. However, the original dispute can still be greatly clarified after the process of justification, and the root of disagreement can be located at the more fundamental intuitions. It is just that we cannot resolve the disagreement itself. I think progress is still being made in the second case as we gain an understanding of our dispute, which might potentially help us resolve the dispute eventually.

Understanding intuitions and substantial disputes in the above way, there are two conditions for convergence to occur concerning any topic or question: first, the majority of the people involved in the discussion share the same relevant intuitions, and second, they understand how these intuitions are connected to the questions of interest. What can we do when people do not share the same relevant intuitions? I think there is very little that we can do, especially when we have intuitions that are almost impossible to argue for or against. The optimistic view here is to consider the differences in people's fundamental intuitions to be less common. It might be the case that although many disputes appear to be caused by differences in intuitions, they are actually caused by verbal misunderstandings and incoherence in reasoning. If this is the case, convergence can be achieved by clarifying meanings and eliminating incoherence.

What about the relationship between intuitions and truths? There is no guarantee that the things we find intuitively true are actually true. This is especially the case with things that people have different intuitions about. If one group of people finds a certain claim intuitively true while the other finds it intuitively false, assuming the claim is truth-apt, one of the groups must be making a mistake. Intuitions here are clearly not reliable guides. At the same time, in

disciplines such as mathematics where people share identical fundamental intuitions, people tend to assume claims or propositions that appear intuitively true to most people to be actually true. There is no guarantee that we are not making mistakes about these claims (which are often called "axioms"), but it is a more than reasonable decision to assume their truth status as it has been working out well for us so far. However, even though intuitions are no guarantee of truths, they are the basis for our beliefs. They are all we have.

So, put simply, to make progress on the zombie argument, and in philosophy in general, we should aim to achieve convergence to the truth on the relevant matters. At the same time, our intuitions can greatly influence whether this goal can be achieved. Progress can still be made through our attempts to clarify and justify the beliefs that we hold. In the process of doing so, we can often point out the more fundamental disagreements in our intuitions that are responsible for our disputes. Sometimes, disputes will be resolved in the process, if the two parties reach an agreement at a more fundamental level. When the disputes cannot be resolved, we can at least locate the root of our disputes by pointing out how our intuitions differ. Disputes caused by differences in unjustifiable fundamental intuitions might be the most difficult ones to resolve, and I do not really have an answer for that. Nonetheless, our attempts to locate our differences in our intuitions and to find justifications for these seemingly unjustifiable beliefs will certainly promote progress in philosophy.

4.2.3 Intuition and the Zombie Argument

The main motivation for me to talk about intuitions comes from my experiences of dealing with disputes over the zombie argument. Although progress on the zombie argument is occasionally made, disputes over the argument sometimes end in stalemates, in which two parties disagree over a claim or proposition but both fail to justify why they adopt their positions. From the analysis earlier, we can say that genuine stalemates are often caused by differences in unjustifiable intuitions, and such differences in intuitions might contribute to the difficulties in resolving the zombie argument. In this section, I will discuss a few disputes from the zombie argument that might be due to the differences in intuitions.

The first case concerns the dispute between the type-A materialists (those who deny the conceivability of zombies) and the zombists. The type-A materialists, especially the analytic functionalists, consider facts about consciousness to be logically entailed by physical facts. As mentioned previously, it is possible to interpret the dispute here as verbal – maybe the analytic functionalists are just using functionally defined concepts of consciousness while the zombists are using non-functionally defined ones. But it is not as simple as that. Once the zombists

clarify the non-functionally defined phenomenal concepts that they are using in the conceivability of zombies, the analytic functionalists can then deny the existence of such concepts or deny that the concepts are capable of referring to anything real. So, it can either be that the analytic functionalists fail to get what we (the zombists) mean by phenomenal consciousness, or that they do get what we mean by the term but still deny that it refers to anything. If they fail to get what we mean, the problem then relates to the extreme difficulty of clarifying the phenomenal concepts. If they do get what we mean, the problem then probably relates to resolving our differences in intuitions.

I do not think there is much we can do to justify either position. If you ask me why I think there is phenomenal consciousness that is not captured by any functionally defined concepts, I will simply tell you that it is obvious and that the existence of phenomenal consciousness is one of the very few things that I am certain of. If you ask an analytic functionalist, you will probably find out that they lack such an intuition. But why? Analytic functionalists do try to provide justifications sometimes. As mentioned in Chapter 2.1, for example, one can argue against the existence of non-functionally defined consciousness by pointing out how its existence is incompatible with our epistemic contact with it (see Balog, 1999; Clark, 2000; Perry, 2001; Kirk, 2005, 2008). By doing so, the dispute of whether phenomenal consciousness exists becomes a combined dispute of whether its existence is incompatible with our epistemic contact with them and whether our epistemic contact with them is important. So, the zombists have the option of providing a theory of content that shows how the existence of phenomenal consciousness causes no issue for epistemic contact, which I believe is what Chalmers (1996, 2003b) has done with the idea of acquaintance. The analytic functionalists then have three options: (1) they can agree with Chalmers's view; (2) they can disagree with the view by simply saying that it is counterintuitive or fails to make sense; and (3) they can disagree with the view by providing a new argument for why they disagree. If the first option is chosen, the dispute is resolved. It has turned out that the analytic functionalists and the zombists do share the same fundamental intuitions. If the second option is chosen, the core of the dispute now lies in the difference in intuitions about the idea of acquaintance. If neither party can provide any further justifications for their positions, a stalemate is reached and the dispute remains unresolved. If the third option is chosen, now it is up to the zombists to decide their moves. This usually ends up with a new dispute over one or multiple premises in the new argument. The cycle repeats until the dispute is resolved or an unresolvable dispute over unjustifiable intuitions appears. In any way, intuitions will be the ones that decide whether the dispute will be resolved and how it will be resolved.

The second case concerns the article Hill and McLaughlin (1999) wrote as an objection to the zombie argument, which also acts as the initial motivation for me to write about intuitions.[52] Put simply, Hill and McLaughlin deny the logical possibility of zombies (premise 2 of the zombie argument) by rejecting the joint exercise of sympathetic and perceptual imagination as a guide to possibility.[53] There are, of course, a lot of moves zombists can make. For example, I can use negative conceivability in premise 2 to avoid the use of imagination, or I can deny the difference between sympathetic and perceptual imagination being genuine. However, the simplest move I can make is just to ask the following: Why is the joint exercise of sympathetic and perceptual imagination not a guide to possibility? I could not find any detailed support provided for this claim by Hill and McLaughlin in their 1999 paper. They do provide some support by saying that Cartesian intuitions are significantly different to stereotypical modal intuitions, but it seems to me that they mostly treat their view as intuitively true. Of course, the view here may still be further justified or argued against, but sometimes both options are difficult and neither party can provide a convincing knockdown argument against their opponent. In cases like this, an interesting problem concerning the burden of explanation arises – which of the two parties should first be demanded to defend their position? Generally speaking, the party that endorses the less intuitive position should be the first to make the move. But again, intuitiveness is subject-related: what appears intuitive to some might appear counterintuitive to others. So, we might say that the party that endorses the view that fewer people find intuitive should carry the burden of explanation. In the case above, I do not know whether more or fewer people find it intuitive that the joint exercise of sympathetic and perceptual imagination fails to be a guide to possibility. So, it is difficult to say which party should have more burdens of justifying their position. I certainly do not find it intuitive here, but if the majority of people do find it intuitive, people who attempt to deny the claim will be obliged to develop an argument explaining why it is the case.

The last specific case I want to discuss here is the case of "Cicero = Tully" by David Papineau (2002). Papineau tries to use this example to break the link between conceivability and possibility. He argues that "Cicero is not Tully" is conceivable but impossible, despite the fact that both "Cicero" and "Tully" refer directly. He then uses this example to show that the transparency thesis, the thesis that "identities involving two directly referring terms are always *a priori* knowable" (Papineau, 2002, p. 92), is false. If the transparency thesis is false, then

[52] Thanks to Peter Godfrey-Smith again for introducing this article to me.
[53] The ideas of sympathetic and perceptual imagination are originally from Nagel (1974).

some directly referring conscious states and directly referring physical states might co-refer without being *a priori* knowable. Papineau's example here can certainly be argued against. The reply should be obvious here: "Cicero" and "Tully" certainly do not refer directly but refer indirectly via contingent features.[54] If the user of these terms simply picks up these terms from other people, the contingent features here are simply *being told by such-and-such person or group of people about the term "Cicero/Tully."* There is certainly another possible world where the user is told about "Cicero" and "Tully" in the same way, but the two terms refer to two different people, assuming that the feature of the two terms co-referring is not already built into the descriptions of the terms. So, if I have the opportunity to write or talk to Papineau, I will explain the above reply to him and hope that he finds it intuitively appealing. But there is also the chance that he fails to find it intuitively appealing. If that is the case, there is not much else I can do. The reply above makes perfect sense to me and appears to clearly demonstrate how "Cicero = Tully" is not a counterexample to the transparency thesis. If Papineau sees it otherwise, he will need to provide an argument or justification for his position. Otherwise, we reach a stalemate due to differences in intuitions.

Once we successfully demonstrate the lack of counterexamples to the transparency thesis, the thesis should not be re-evaluated. I certainly find the transparency thesis intuitively appealing, even though there is not much I can say to justify it besides explaining why all the proposed counterexamples fail. However, the lack of counterexamples does not automatically make the thesis true. Papineau, and maybe some other philosophers, can still consider the thesis to be false by saying something along the lines of "I see no good reason to support such a thesis." If there is a convincing argument coming from either party, the dispute will be resolved. Otherwise, we will have another stalemate.

It is not my goal to provide a detailed analysis of any of the cases above here due to the word limit of this thesis. However, I do hope to use the above cases to shed some light on the relationship between disputes and intuitions. I do not have any answer to how stalemates can be dealt with, but I still hope the analysis above more or less helps us understand why certain disputes are so difficult to resolve.

So far, I have painted quite a pessimistic picture, saying that some disputes are potentially unresolvable due to differences in fundamental intuitions. I think that we should still be optimistic. I think at least a large number of substantial disputes are resolvable. The method of justification mentioned earlier will at least help us locate the source of the dispute

[54] Braddon-Mitchell (2004) also argues for a similar idea.

and how we differ in our intuitions. Sometimes, our differences in certain intuitions can be overruled if we turn out to share the more fundamental intuitions. We simply need to discover these fundamental intuitions and demonstrate how they can be used to construct satisfactory answers to our big questions in philosophy. Although there might be unresolvable disputes due to differences in some fundamental intuitions, I believe there is still plenty of progress to be made towards the zombie argument.

Concluding Remarks

In this thesis, I have set out to provide a thorough analysis and further defence of the two-dimensional zombie argument. Although I do not have the intention to settle the debate once and for all, I do hope to at least demonstrate that the zombie argument is in a highly defensible position once it is understood properly. To achieve this goal, I first clarified the argument by analysing the core concepts involved in the argument in Chapter 1, which aims to eliminate any potential misunderstandings of the argument. In Chapter 2, I discussed the objections and defences made towards the argument and demonstrated that the argument is highly defensible. Chapter 3 is the place where I contributed to the debate by defending the argument against objections from three recent articles. Finally, in Chapter 4, I speculated on the root of disagreements over the argument by analysing how the use of language and intuitions might generate disputes.

Overall, there do not appear to be any knockdown objections to the zombie argument. Many objections to the basic version of the zombie argument lose their force against the two-dimensional version, and some objections specifically designed to target the two-dimensional argument seem to have misunderstandings of certain ideas involved in the argument which render the objections ineffective. So, for starters, I think anyone intending to formulate objections to the two-dimensional zombie argument should first be familiar with the relevant ideas involved such as two-dimensional semantics, the o-physical conception, phenomenal consciousness, conceivability, and Russellian monism. Otherwise, any objection formulated will be missing the targets and failing to pose any genuine threat. There are certainly some objections that get the argument right and raise genuine concerns about the argument, but most, if not all, of these objections rely on more or less debatable assumptions. For example, most objections related to zombies' phenomenal judgements depend on some specific theories of reference and content, and very often, these background theories are topics of heated debates themselves. I am not saying that these assumptions, or the objections themselves, are definitely false. However, we should take these objections with a grain of salt. At the very least, there is no notable objection that defeats the two-dimensional zombie argument in a convincing way.

Besides clarifying the argument and summarizing the previous discussions on the argument, I also provided detailed defences against objections featured in three articles: one by Phillip Goff and David Papineau (2014); one by Daniel Stoljar (2020); and one by Eugen Fischer and Justin Sytsma (2021). These defences also count as the contribution I am trying to make to the debate. I do believe that the two-dimensional zombie argument is sound, but even if it is in fact sound, convincing everyone of its soundness will no doubt be a difficult task. The

best way to make progress, I think, is to keep the debate going. Potential objections to the argument should always be welcome, but at the same time, defences should also be actively made. The more objections the argument survives, the more believable the argument becomes. On the other hand, if one or multiple objections are shown to be resilient, the argument naturally loses its strength.

If the two-dimensional zombie argument is sound, we can conclude that structural physicalism is false and that either dualism or Russellian monism is true. Whether dualism or Russellian monism is true depends on whether physical concepts and phenomenal concepts co-refer. It should be emphasized again that such co-reference does not automatically grant the truth of physicalism, as for such co-reference to occur, the physical concepts involved must be capable of indirectly referring to quiddities (for quiddities, see Chapter 1.7). The resulting view will be dramatically different to the traditional physicalism most philosophers are familiar with. Although the argument here is denying physics' ability to fully reveal the nature of consciousness, it is not saying that the physical science tells us nothing about consciousness. With some reasonable assumptions, we can still make plausible inferences about consciousness by studying physical phenomena. The view here is by no means anti-scientific. It is just that physics does not tell us everything about consciousness.

Last but not least, I think there is still room for progress to be made on the debate. As I have mentioned earlier, both objections and defences of the argument should be welcome. When disagreements do arise, the two parties should focus on clarifying their use of language and justifying their use of intuitions. One concept that philosophers might want to focus on clarifying, I think, is the concept of *physical*. Although we have an intuitive understanding of certain things and properties being physical, I do not think there is yet a satisfactory conception to thoroughly explain what it means to be physical. Coming up with such a conception might greatly influence how we evaluate the zombie argument and open up opportunities to settle the debate on the argument once and for all.

References

Alter, T., & Nagasawa, Y. (2012). What is Russellian monism?. *Journal of Consciousness Studies*, *19*(9-10), 67-95.

Armstrong, D. M. (1968). A Materialist theory of mind.

Bailey, A. (2007). The unsoundness of arguments from conceivability. *University of Guelph. Ontario, Canada*.

Balog, K. (1999). Conceivability, possibility, and the mind-body problem. *Philosophical Review*, 497-528.

Blackburn, S. (1990). Filling in space. *Analysis*, *50*(2), 62-65.

Block, N. (1980a). Are absent qualia impossible?. *The Philosophical Review*, *89*(2), 257-274.

Block, N. (1980b). Troubles with functionalism. In *The Language and Thought Series* (pp. 268-306). Harvard University Press.

Block, N., & Stalnaker, R. (1999). Conceptual analysis, dualism, and the explanatory gap. *The Philosophical Review*, *108*(1), 1-46.

Bourget, D., & Chalmers, D. J. (2014). What do philosophers believe?. *Philosophical studies*, *170*, 465-500.

Braddon-Mitchell, D. (2003). Qualia and analytical conditionals. *The Journal of Philosophy*, *100*(3), 111-135.

Braddon-Mitchell, D. (2004). Masters of our meanings. *Philosophical Studies*, *118*, 133-152.

Braddon-Mitchell, D., & Jackson, F. (1996). Philosophy of mind and cognition: An introduction.

Braddon-Mitchell, D., & Jackson, F. (1999). The Divide and Conquer Path to Analytical Functionalism. *Philosophical Topics*, *26*(1/2), 71-88.

Campbell, K. (1970). *Body and mind*.

Chalmers, D. J. (1996). *The conscious mind: In search of a fundamental theory*. Oxford Paperbacks.

Chalmers, D. J. (1999). Materialism and the Metaphysics of Modality.

Chalmers, D. J. (2002). Does conceivability entail possibility?. *Conceivability and possibility*, *145*, 200.

Chalmers, D. J. (2003a). Consciousness and its place in nature. *Blackwell guide to the philosophy of mind*, 102-142.

Chalmers, D. J. (2003b). The content and epistemology of phenomenal

Chalmers, D. J. (2004a). Imagination, indexicality, and intensions.

Chalmers, D. J. (2004b). Epistemic two-dimensional semantics. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, *118*(1/2), 153-226.

Chalmers, D. J. (2006a). Phenomenal concepts and the explanatory gap. In (T. Alter & S. Walter, eds) Phenomenal Concepts and Phenomenal Knowledge: New Essays on Consciousness and Physicalism.

Chalmers, D. J. (2006b). The foundations of two-dimensional semantics. *Two-dimensional semantics*, 55-140.

Chalmers, D. J. (2006c). Two-Dimensional Semantics. In *Oxford Handbook of Philosophy of Language, E. Lepore and B. Smith (eds.), Oxford: Oxford University Press*, pp. 575–606.

Chalmers, D. J. (2009). The two-dimensional argument against materialism.

Chalmers, D. J. (2011). Verbal disputes. *Philosophical Review*, *120*(4), 515-566.

Chalmers, D. J. (2012). *Constructing the world*. OUP Oxford.

Chalmers, D. J. (2014). Intuitions in philosophy: A minimal defense. *Philosophical Studies*, *171*(3), 535-544.

Chalmers, D. J. (2015a). Panpsychism and panprotopsychism. *Consciousness in the physical world: Perspectives on Russellian monism*, *246*.

Chalmers, D. J. (2015b). Why Isn't There More Progress in Philosophy? 1. *Philosophy*, *90*(1), 3-31.

Chalmers, D. J. (2020). Spatiotemporal functionalism v. the conceivability of zombies. Noûs, 54(2), 488-497.

Chalmers, D. J., & Jackson, F. (2001). Conceptual analysis and reductive explanation. *The Philosophical Review*, *110*(3), 315-360.

Churchland, P. S. (1996). The hornswoggle problem. *Journal of Consciousness Studies*, *3*(5-6), 402-408.

Clark, A. (2000). A case where access implies qualia?. *Analysis*, *60*(1), 30-37.

Crane, T., & Mellor, D.H. (1990) There is no Question of Physicalism. *Mind, 99*: 185–206.

Dennett, D. C. (1993). *Consciousness explained*. Penguin uk.

Dennett, D. C. (1996). Facing backwards on the problem of consciousness. *Journal of Consciousness Studies*, *3*, 4-6.

Dowell, J. L. (2006). Formulating the thesis of physicalism: an introduction. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, *131*(1), 1-23.

Engel, A. K., & Singer, W. (2001). Temporal binding and the neural correlates of sensory awareness. *Trends in cognitive sciences*, *5*(1), 16-25.

Evans, G. (1979). Reference and contingency. *The Monist*, *62*(2), 161-189.

Fischer, E., & Sytsma, J. (2021). Zombie intuitions. *Cognition*, *215*, 104807.

Frankish, K. (2007). The Anti-Zombie Argument. *The Philosophical Quarterly*, *57*(229), 650-666.

Gettier, E. L. (1963). Is Justified True Belief Knowledge?. *Analysis*, *23*(6), 121-123.

Goff, P. (2011). A posteriori physicalists get our phenomenal concepts wrong. *Australasian Journal of Philosophy*, *89*(2), 191-209.

Goff, P. (2017). *Consciousness and fundamental reality*. Oxford University Press.

Goff, P., & Papineau, D. (2014). What's wrong with strong necessities?. *Philosophical Studies*, *167*(3), 749-762.

Hempel, C. (1969). Reduction: Ontological and linguistic facets. In *S. Morgenbesser, et al. (eds.), Essays in Honor of Ernest Nagel, New York: St Martin's Press*, pp. 179–199.

Hempel, C. (1980). Comments on Goodman's Ways of Worldmaking. *Synthese* 45: 193 - 199.

Hill, C. S., & McLaughlin, B. P. (1999). There are fewer things in reality than are dreamt of in Chalmers's philosophy.

Jackson, F. (1982). Epiphenomenal Qualia. *The Philosophical Quarterly*, *32*(127), 127-136.

Jackson, F. (1998). *From metaphysics to ethics: A defence of conceptual analysis*. Clarendon Press.

Kim, J. (1984). Concepts of supervenience. *Philosophy and phenomenological research*, *45*(2), 153-176.

Kirk, R. (1974). Sentience and behaviour. *Mind*, *83*(329), 43-60.

Kirk, R. (2005). *Zombies and consciousness*. Clarendon Press.

Kirk, R. (2008). The inconceivability of zombies. *Philosophical Studies*, *139*(1), 73-89.

Kirk, R. (2019). Zombies. *Stanford Encyclopedia of Philosophy*.

Koch, C., Massimini, M., Boly, M., & Tononi, G. (2016). Neural correlates of consciousness: progress and problems. *Nature Reviews Neuroscience*, *17*(5), 307-321.

Kripke, S. A. (1972). Naming and necessity. In *Semantics of natural language* (pp. 253-355). Springer, Dordrecht.

Lamme, V. A. (2006). Towards a true neural stance on consciousness. *Trends in cognitive sciences*, *10*(11), 494-501.

Laurence, S., & Margolis, E. (1999). Concepts and cognitive science.

Levine, J. (1983). Materialism and qualia: The explanatory gap. *Pacific philosophical quarterly*, *64*(4), 354-361.

Levine, J. (2020). A Posteriori Physicalism: Type-B Materialism and the Explanatory Gap. *The Oxford Handbook of the Philosophy of Consciousness*, 387-404.

Lewis, D. K. (1966). An argument for the identity theory. *The Journal of Philosophy*, *63*(1), 17-25.

Lewis, D. K. (1970). How to define theoretical terms. *The Journal of Philosophy*, *67*(13), 427-446.

Lewis, D. K. (1994). Reduction of mind. In (S. Guttenplan, ed.) *A Companion to the Philosophy of Mind*. Oxford:Blackwell.

Lewis, D. K. (1995). Should a materialist believe in qualia?.

Lewis, D. K. (2001). *Ramseyan humility*. Department of Philosophy, University of Melbourne.

Loar, B. (1990). Phenomenal states. *Philosophical perspectives*, *4*, 81-108.

Loar, B. (1999). David Chalmers's The Conscious Mind.

Lockwood, M. (1989). *Mind, brain and the quantum: The compound 'I.'*. Basil Blackwell.

Ludwig, K. (2003). The mind-body problem: An overview. The Blackwell guide to philosophy of mind, 1-46.

Lycan, W. G. (1996). *Consciousness and experience*. Mit Press.

Marcus, E. (2004). Why zombies are inconceivable. *Australasian Journal of Philosophy*, *82*(3), 477-490.

Maxwell, G. (1978). Rigid designators and mind-brain identity.

McGinn, C. (1989). Can we solve the Mind--Body problem?. *Mind*, *98*(391), 349-366.

Melnyk, A. (2003). Physicalism. *The Blackwell guide to philosophy of mind*, 65-84.

Montero, B., & Papineau, D. (2005). A defence of the via negativa argument for physicalism. *Analysis*, *65*(3), 233-237.

Nagel, T. (1974). What is it like to be a bat?. *The philosophical review*, *83*(4), 435-450.

Ney, A. (2008). Defining physicalism. *Philosophy Compass*, *3*(5), 1033-1048.

Papineau, D. (2002). *Thinking about consciousness*. Clarendon Press.

Papineau, D. (2006). Phenomenal and Perceptual Concepts. *Phenomenal concepts and phenomenal knowledge: New essays on consciousness and physicalism*.

Perry, J. (2001). *Knowledge, possibility, and consciousness*. mit Press.

Piccinini, G. (2017). Access denied to zombies. *Topoi*, *36*(1), 81-93.

Prelević, D. (2017). Access granted to Zombies. *Theoria, Beograd*, *60*(1), 58-68.

Putnam, H. (1975). The meaning of ‘meaning’. *Philosophical papers*, *2*.

Quine, W. V. (1951). Two Dogmas of Empiricism, 60 Phil. *Rev*, *20*, 20-34.

Rees, G., Kreiman, G., & Koch, C. (2002). Neural correlates of consciousness in humans. *Nature Reviews Neuroscience*, *3*(4), 261-270.

Russell, B. (1927). The Analysis of Matter.

Shoemaker, S. (1975). Functionalism and qualia. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, *27*(5), 291-315.

Shoemaker, S. (1999). On David Chalmers's the conscious mind.

Smart, J. J. (1959). Sensations and brain processes. *The Philosophical Review*, *68*(2), 141-156.

Spurrett, D., & Papineau, D. (1999). A note on the completeness of 'physics'. *Analysis*, *59*(1), 25-29.

Stoljar, D. (2001a). Two conceptions of the physical. *Philosophy and Phenomenological Research*, *62*(2), 253-281.

Stoljar, D. (2001b). The conceivability argument and two conceptions of the physical. *Philosophical Perspectives*, *15*, 393-413.

Stoljar, D. (2005). Physicalism and phenomenal concepts. *Mind & language*, *20*(5), 469-494.

Stoljar, D. (2010). *Physicalism*. Routledge.

Stoljar, D. (2020). Chalmers v Chalmers. *Noûs*, *54*(2), 469-487.

Stoljar, D. (2022) Physicalism. *Stanford Encyclopedia of Philosophy*.

Thomas, N. J. (1998). Zombie killer. *Toward a science of consciousness II*, 171-77.

Tiehen, J. (2018). Physicalism. *Analysis*, *78*(3), 537-551.

Van Gulick, R. (1993). Understanding the Phenomenal Mind: Are We All Just Armadillos?. In M. Davies and G. Humphreys (eds.), Consciousness: Philosophical and Psychological Aspects. Oxford: Blackwell.

Van Gulick, R. (2014). Consciousness. *Stanford Encyclopedia of Philosophy*.

Worley, S. (2003). Conceivability, possibility and physicalism. *Analysis*, *63*(1), 15-23.

Yablo, S. (1993). Is conceivability a guide to possibility?. *Philosophy and Phenomenological Research*, *53*(1), 1-42.

Yablo, S. (1999). Concepts and consciousness. *Philosophy and Phenomenological Research*, *59*(2), 455-463.