This work is protected by copyright and other intellectual property rights and duplication or sale of all or part is not permitted, except that material may be duplicated by you for research, private study, criticism/review or educational purposes. Electronic or print copies are for your own personal, non-commercial use and shall not be passed to any other individual. No quotation may be published without proper acknowledgement. For any other use, or to quote extensively from the work, permission must be obtained from the copyright holder/s.

The consciousness science paradox

Adam Lee Balmer

Doctor of Philosophy in Philosophy

June 2019

Keele University

Abstract

This thesis argues that there are two claims we can persuasively make about consciousness. The first is that we should be able to identify consciousness through the sort of empirical observation characteristic of standard scientific practices. The second is that we cannot identify consciousness through empirical means. Both of these claims are defended, with the first being established by reference largely to the problems of accounting for our capacity to reliably report our conscious experiences, and the latter being established by showing how scientific evidence cannot be used to tell us the truth or falsehood of claims about consciousness without making *a priori* assumptions that undermine the very prospect of a scientific approach to the study of consciousness. The paradox arising as a result of these apparently contradicting claims is shown not to be a consequence of a particular perspective in philosophy of science or philosophy of mind as, when a variety of perspectives in both areas are surveyed, none are shown to avoid the problem without running into similar difficulties. I conclude by providing a diagnosis of the philosophical roots of the paradox.

Contents

Acknowledgements

Special thanks to my brother Kyle, Mum and Dad. Without your support, I don't know how this thesis would have been possible. My work was also helped along by my grandma, who, among other things, has always been willing to supply me with tea and listen to me rant, and by my girlfriend Nikita Ganser, who has supported me in lots of ways over the last couple of years. I also owe thanks to my friend Elizabeth Nelson for helping my do sentences good and to all the other friends whose discussions have helped develop my ideas over the years (Robert Reid, Steve Hyde, Gareth Day, Christopher Murphy, George Carpenter, Sebastian Orlander and others). And finally, a big thank you to my supervisor James Tartaglia for patiently steering me toward writing something I can be proud of.

Introduction

To look at the wildly disparate philosophical views on the nature of consciousness, you could be forgiven for assuming that philosophers are intensely confused about it. The stances vary so widely that the view that consciousness may not exist has been defended (e.g. James 1904a), and so has the view that everything exists within consciousness (e.g. Sprigge 1983; Bolender 2001). There are those who have argued that we are cognitively incapable of understanding consciousness (McGinn 1993: 27–45) and those who have argued that many seemingly major problems in understanding consciousness are artificial, resulting from reflections used in philosophical discourse leading us easily into confusion (Dennett 2013: 279–353). Lockwood stated in 1993 that there was "no glimmer of a consensus amongst philosophers about the mind-body problem" (Lockwood 1993: 271) and this situation does not appear to have improved.

The philosophical confusion is interesting given that there is a strong intuitive sense that consciousness is what we are most familiar with. Our consciousness is what is present whenever we perceive the world from our own perspective, when we are presented with our own thoughts, emotions and feelings. That consciousness, which is present to us continuously throughout our waking lives, should invite such diverse philosophical speculation is almost as intriguing as the puzzles themselves.

Philosophical confusion can also be contrasted with the apparent absence of such confusion in scientific discourse. Scientists claim to be studying consciousness and use common research methods without spending anywhere near the same quantities of time and effort filling their papers and books with debate about which metaphysical picture their results are to be interpreted under.

We could interpret this relative scientific consensus with regard to the sort of approach that is required to understand consciousness as a vindication of the view that philosophers are

simply indulging in largely semantic issues or taking too seriously thought experiments about zombies (Chalmers 1996: 94–100), inverted colour spectra (ibid.: 99–101) and the potential experiences of electrons (Deiss 2009: 153). This would be to assume, though, that scientists make no philosophical assumptions, or that they only make correct ones, both of which are demonstrably false. As Dennett stated, "there is no such thing as philosophy-free science; there is only science whose philosophical baggage is taken on board without examination" (Dennett 1995a: 21).

It cannot both be true that consciousness, forming a part of the world just like the many objects of successful empirical investigations, should seamlessly enter into our scientific discourse, while it is also true that this scientific discourse is incapable of making any progress with regard to our understanding of consciousness. I will argue for both of the above claims in this thesis.

This second claim may appear to be surprising. Even many of those philosophers who regard consciousness as, in an important sense, lying outside the scope of scientific investigation, consider it to be possible to at least use scientific means to draw up correlations between the presence of consciousness and other empirically observable states, such as neurophysiological states (Chalmers 2003: 1113). This task, I wish to show, has not only failed utterly so far, but is *in principle* impossible to ever achieve.

If indeed such a paradox can be demonstrated, this will present a significant philosophical challenge for all philosophical perspectives for which the two incompatible truths hold, both that consciousness should and should not be a proper object of scientific enquiry. As it will be shown, this range of perspectives is surprisingly wide, and related paradoxical results arise for perspectives that deny the very existence of consciousness.

I will now outline in brief the arguments that are made to get to this conclusion. We first require a characterisation of consciousness that does not presuppose the truth of any specific metaphysical perspective, since the paradox I want to bring out is not specific to a particular viewpoint on the nature of consciousness. We arrive at such a characterisation in

chapter 1. We establish this and ascertain that the most plausible starting point for any investigation into consciousness is to study those states that we can behaviourally indicate the presence of, which I describe as "paradigmatic conscious states" [1.1]. We then consider how to defend this characterisation from criticism by those espousing epiphenomenalism [1.2], with focus on Chalmers' philosopher's zombie argument [1.2.1] and Jackson's knowledge argument [1.2.2], before evaluating a common response to the knowledge argument to show that the response that knowledge-by-acquaintance is possible gives us no means of supposing that paradigmatic conscious states could be causally inefficacious [1.2.3]. We then assess the argument that the causal efficacy of such paradigmatic conscious states entails that they should be possible to empirically identify [1.3].

In chapter 2, we assess what constitutes an attempt to study consciousness scientifically and show some ambiguities related to the task of such research. It is shown that for something to be studied scientifically it must be possible to use empirical observation to confirm or refute claims about that thing [2.1], and how this is compatible with Feyerabend's anarchism [2.1.1] and contrary to Ladyman, Ross and their contributors' institutionalism, but that this latter view gives us a flawed understanding of science anyway [2.1.2]. In [2.2], we see that attempts over the last few decades to understand consciousness, as epitomised in the work of Bernard Baars [2.2.1] and Francis Crick [2.2.2], rely on a contrastive approach that distinguishes between the presence and absence of paradigmatic states of consciousness. This leads to the conclusion that the task of consciousness science is to identify those states that are empirically observable correlates of consciousness [2.3], which can only be done if we distinguish the presence of paradigmatic states of consciousness from their absence based on the presence of a capacity to produce certain behaviour, notably reports of conscious experience [2.4]. Finally, we ascertain that there is no way to demarcate a state that has this capacity and one that lacks it, meaning that the evidence that the whole scientific project hinges on is fundamentally flawed [2.5].

In chapter 3, we determine that, even if we only wish to study paradigmatic states of consciousness, this does not entail that the conscious states in question cannot be separated from their capacity to produce behavioural indicators of their presence [3.1], and we assess possible cases where the consciousnesses do lack this capacity [3.2]. We see that this possibility would doom any attempt to study consciousness scientifically to failure [3.2.1], and that we cannot simply deny that such a thing is possible in principle without similarly dooming a science of consciousness to failure [3.2.2]. We then examine the claim that it is unlikely that such states exist and find it to be similarly flawed [3.2.3] before establishing that it is unusual for a scientific approach to be so susceptible to the results of thought experiments [3.2.4]. We then examine the possibility that a consciousness science can distinguish between the presence and absence of particular paradigmatic conscious states [3.3]. We observe that this relies on the assumption that we know when a conscious state ends and when it begins, which requires us to know that there is a singular persisting consciousness at a time for each person [3.3.1]. We examine some arguments against the possibility that a consciousness could divide or that multiple consciousnesses could combine within a person [3.3.2] before ascertaining that ruling out the possibility *a priori* would render a science of consciousness impossible [3.3.3]. We then establish that there are no verificationist grounds or reasons concerning probability for assuming such a thing to be implausible [3.3.4]. This leads us to the conclusion that there is no way to empirically observe the correlates of paradigmatic conscious states [3.4].

In chapter 4, we question whether it is possible at least to establish when a conscious state occurs in relation to other empirically observable events. We use Libet et al.'s famous experiments on subjective delay as the basis for the discussion [4.1], and are presented with arguments that the conclusions the researchers arrive at are philosophically problematic on multiple grounds, and most relevantly that they assume that it is possible to independently observe what is taking place in consciousness and what is taking place independently of consciousness. We examine several positions that may be thought to give a philosophical

grounding to this assumption, such as transparency [4.2], the idea that perceptual states are produced by the presence of the object of perception [4.3], the notion that we derive our concept of time from the succession of conscious events [4.4] and the structural resemblance between conscious events and mind-independent events [4.5], but find that none of these positions gives us reason to suppose that we can ascertain the timing of conscious events related to other empirically observable events. We arrive at the conclusion that there can be no science of consciousness as currently envisaged [4.6].

In chapter 5, we are introduced to the Consciousness Science Paradox. The two contradicting claims have now been established and are highlighted [5.1], namely that paradigmatic states of consciousness both must be possible to study scientifically and that they cannot be studied scientifically. We see that consciousness scientists are actually studying certain behavioural and cognitive capacities [5.2], and that there is no need for them to refer to these as states of consciousness at all. It is shown that this is a problem either for our conception of science or our conception of consciousness, if not both [5.3]. We thus begin the attempt at evaluating our conception of science [5.4] by first reassessing our view of science to see if we can conceive of it in a way that would permit the possibility that consciousness is observable [5.5]. We examine Popper's falsificationism [5.5.1], Kuhn's view of scientific revolutions [5.5.2] and Bayes' theorem [5.5.3] in order to arrive at the conclusion that the paradoxical result persists through all of these approaches [5.5.4]. We thus attempt to conceive of science in a manner such that any requirement for science to explain consciousness is removed [5.6], using Lakatos' research programmes [5.6.1] and Feyerabend's anarchism [5.6.2] as such examples, before arriving at the conclusion that even if we deny the requirement for science to examine consciousness this still fails to give us a scientific study of consciousness [5.6.3]. We finally examine the possibility that philosophically unaided scientific progress will yield a description of consciousness that is not so problematic [5.7]. We look at two different interpretations of wave-function collapse in quantum mechanics to demonstrate that any interpretation of an observable phenomena that assumes the

phenomena are dependent upon those phenomena being consciously observed is doomed to failure.

Turning away from modifying our conception of science, we look to the possibility that we can modify our conception of consciousness to avoid the paradox in chapter 6. We look at three examples, which are physicalism [6.1], property dualism [6.2] and Russellian monism [6.3], and find that all of these perspectives require that consciousness must have specifiable observable effects, otherwise they collapse into the epiphenomenalism we have already rejected in [1.2], and yet the fact that such observable effects must exist and yet cannot be is exactly what produces the paradox [6.4].

In chapter 7, we attempt to deny the existence of consciousness to avoid the paradox. We first attempt to give a characterisation of the neutral monism of Mach, James and Russell by comparing it to Coleman's panqualityism [7.1], where we ascertain that panqualityism lacks the neutral aspect of neutral monism, which would require us to deny the existence of metaphysically substantive entities such as subjects or consciousnesses. Moreover, still maintaining that conscious states correspond to other empirically observable states, panqualityism is found to be subject to the same paradoxical result as the other aforementioned positions. The difficulties of neutral monism are then ascertained [7.2] whereby we establish that certain neutral monist positions, such as Russell and Mach's, rely on there being a correlation between neurophysiological states and percepts that is equally as impossible to establish through empirical observation, whereas James' radical empiricism relies on there being a succession of experiences that, as established in chapter 4, cannot be established by empirical observation. We then examine eliminative materialism in some depth [7.3] and establish that it lies between two difficulties in that it must choose between undermining its own epistemic basis by denying that empirical observation is possible or it must assume that neurophysiological states can underpin states of perceptual awareness, in which case a similar paradoxical result occurs. We examine illusionism, only to find that

replacing talk of phenomenal properties with talk of phenomenal judgements produces the same paradoxical result [7.4].

In chapter 8, we begin the project of diagnosing the source of the Consciousness Science Paradox. We begin by attempting to give an account of what would constitute an observation of a conscious state, determining that the very idea is problematic [8.1]. We attempt to avoid these problems by theorising that direct objects of perception could be publicly observable in adopting a naive realist perspective, but the disjunctive thesis that typically goes along with this ends up with a similar paradoxical result but with reference to how it seems to be in a certain conscious state rather than actually being in that state [8.2]. When we return to the denial of the existence of consciousness, we find that the same paradoxical result is produced but with reference to our capacity to judge ourselves to be in a certain conscious state, and so we settle on a diagnosis of the paradox as arising given any judgement of ourselves to be in a certain conscious state [8.3].

This completes our introduction to and preliminary investigation of the Consciousness Science Paradox.

1

The Causal Efficacy of Consciousness

[1.1] Characterising consciousness

To begin any sort of inquiry about consciousness, we will need to have some sort of operative definition that we can use. I will define consciousness and conscious states as follows:

Consciousness: The capacity for a being to have a subjective perspective on the world.

Conscious state: An instance of subjective perspective.

The term “subjective perspective” refers to an individual’s point-of-view with regards to the world. In perceiving the world, there is a certain way it appears to me with things appearing at certain distances away, visually appearing at a particular angle, with only certain aspects of many things being perceptible at all, and there are certain things I am aware of that seem directly related to my present condition, such as the feeling of the chair I am sitting on and the temperature of the air around me.

Even those things that I am aware of but that do not have any easily specifiable spatial relation to my present condition fall within my overall perspective. If I feel joy, there may be nothing about this joy that is directly related to any specific element of my surroundings or even obviously related to any part of my bodily condition but the fact that I am, at that moment,

aware of my feeling of joy rather than a feeling of despair places this feeling within the overall field of awareness I am referring to as a "subjective perspective."

Whether subjective perspective necessarily involves qualitative properties will not be specified here. What it means for a property to be "qualitative" in this sense is often described demonstratively; the qualitative properties involved in vision are how a particular scene *looks to you*, with all the distinctive hues and shapes that form the appearance of what you are looking at. Described as constituting the way things appear to us rather than the way they actually are, qualitative properties are often labelled "phenomenal properties".

Often phenomenal properties associated with vision are described with particular reference to distinct qualities that are taken to be experiential by nature, with the redness that we are aware of in seeing a ripe tomato being a distinct phenomenal property. Another term used in the description of the properties of consciousness is "qualia", which can be described as properties that are inherently qualitative by nature. Many philosophers argue that consciousness cannot be properly described or explained without such qualitative properties being included in the description. Nagel famously described the major philosophical difficulty of consciousness as being that we have no idea how to describe *what it is like* to be in a particular conscious state other than by simply being in it (Nagel 1974), and Chalmers proposed that the really *hard* problem of consciousness is in explaining why there is a subjective character to our experience at all (Chalmers 1995). Consciousness thus characterised, as being constituted of or containing properties with private, qualitative characteristics, has been referred to as "phenomenal consciousness" (e.g. Tye 2000: ix).

Numerous philosophers have denied the existence of qualitative properties, including some reductive physicalists (Smart 1959; Place 1956) and the extent to which one can coherently maintain the existence of phenomenal qualities and be a reductive physicalist has been questioned (Tartaglia 2013). Dennett has denied the existence of qualitative properties while still accepting that consciousness exists, citing a detailed example of having a particular subjective perspective as he sat in his rocking chair reading a book, looking outside through a

wrinkle in his window and noticing that his rocking synchronised with a musical piece by Vivaldi playing in the background, his thoughts and his pleasure in thinking them (Dennett 1991a: 26–27). Denying that qualitative properties exist does not entail believing that subjective perspective cannot be explained. In any case, it is possible to deny the existence of qualitative properties but accept that the term "subjective" has some meaning otherwise. Russell distinguished between different kinds of subjectivity, such as "physical subjectivity" where the medium responsible for sense perception (i.e. light waves) is affected by something in the environment in the way that a stick is made to look bent by being half in water, and "physiological subjectivity" arising from defects in sense organs (Russell 1921: 225). These forms of subjectivity do not need to involve intrinsic qualities, but only for the object receiving information to have a certain physical status in relation to the things it is "observing". Thus, for the philosopher who denies qualitative properties, it is possible to give an account of subjectivity that does not invoke any intrinsic, phenomenal qualities.

I do not wish to make any attempt to resolve the dispute over whether or not consciousness must be constituted of qualitative properties at this stage, nor indeed will my focus be on the problem of why (if at all) qualitative properties accompany certain physical states, despite Chalmers' assertion that "if any problem qualifies as the problem of consciousness, it is this one" (Chalmers 1995: 5). My focus will be on the incongruity between the facts that consciousness seems to be a candidate for empirical investigation and that it fails to be such. With regard to this latter fact, I will attack the notion, widely accepted on both sides of the divide on the existence or non-existence of phenomenal qualities, that we have a basic understanding of consciousness such that we are able to discern in which circumstances a conscious state is present and in which circumstances it is not. The situation, I will argue, is direr than that and we lack even a rudimentary understanding of the circumstances under which conscious states occur.

There are other variations on our understanding of what constitutes an instance of subjective awareness that depend on our metaphysical position. If I am a panpsychist,

accepting that all physical things “have mind, or instantiate mind, or embody mental states” (Skrbina 2009: xii), then I will judge all physical states to be conscious states. If I am a mindbrain identity theorist, I will regard all conscious states to be particular brain states. To take an example, we might judge that a waking person walking around and interacting with their environment is in a particular conscious state, but we might not make such a judgement about an individual with epileptic automatism having an “absence seizure.” Such individuals behave in a superficially similar way to people who are not undergoing an absence seizure, but they lack any signs of self-awareness as well as any capacity to recall these periods (Damasio 2010: 162–164). For the identity theorist, there is no problem in stating that a person in an ordinary state of awareness is conscious, but that the person having an absence seizure is not; all that needs to be the case for such a statement to be true is that the physiological state of the epileptic individual at that moment is, in some particular way, different to that of the same individual when they are not having a seizure. For the panpsychist, on the other hand, there is consciousness present in both cases, even if that consciousness is qualitatively distinct and fails to produce the same physiological and behavioural symptoms.

As such, to avoid hopeless ambiguity as to whether we are referring to conscious states or are misappropriately labelling certain states as conscious, the scope must be narrowed to the conscious states that we are familiar with in our everyday experience. I assume that all positions that describe consciousness would agree that conscious states are present in cases where I am able to behave in a way that we are familiar with being accompanied by conscious states, such as if I am able to verbally report that I am feeling pain, or that I can smell something sweet. Rather than focussing on the ambiguity that arises in other examples where no such behaviour is present, we should focus our analysis on the points where all perspectives meet, which is with regard to subjective perspectives that we are able to communicate. These are also the states that seem most amenable to empirical study, since we indicate their presence through behaviour. It is thus paradigmatic states of consciousness that will be the object of investigation in this thesis:

Paradigmatic conscious states (hereafter PC states): Conscious states that an individual can indicate the presence of through behaviour, such as reports.

Provided we are able to have some basic understanding of PC states (which I will argue that we are not), it should be much easier to pinpoint in which situations we can find such states than it should be for currently non-confirmable states of consciousness, since the presence and nature of the latter is much easier to dispute. By "currently non-confirmable states of consciousness", I refer to states such as those that a panpsychist would refer to in positing that, say, an electron was to be in or contain states of consciousness, since we would presumably not be privy to such a subjective perspective. By giving what I take to be as close to a basic and uncontroversial notion of conscious states as I can arrive at, this should provide a basis upon which I can criticise current approaches; if even PC states cannot be understood on a rudimentary level such that we are able to describe in which situations they are present, this will serve as a problem for all metaphysical views so far discussed and many others. Yet there are two threats to such a definition of PC states that I will need to contend with first: ineffability and epiphenomenalism.

If conscious states are indeed ineffable, we may think that states of awareness cannot really be communicated. As such this would pose a problem for the above description of PC states since we might take it that no states suffice to enable an individual to communicate their subjective perspective. Yet, the ineffability of conscious states is based on the supposed difficulties of knowing what it is like to be in a conscious state if you have never been in a similar state before, such as knowing what it feels like to use echolocation as a means of finding your way around if you are not an animal who uses echolocation, such as a bat (Nagel 1974). The same problem does not arise between two organisms having similar conscious states and discussing them, such as if I am to describe my feeling tired to another human who

has been in a similar state before. While I am only communicating a state that the other person is familiar with, that familiarity, according to ineffability views, is simply a precondition for such communication, not something that precludes communication. I should also add that it is not clear to me that the same problem does not apply to discussion that requires any concepts at all and not simply to concepts related to conscious experience; I can discuss mathematics with somebody only to the extent that we both have a common understanding of mathematical concepts. If somebody fails to understand what I mean in relating a mathematical fact to them, then the only way I can help them to understand is to guide them using concepts that I have reason to think they already have some grasp of, such as by analogy with something else or by combining simpler or more commonly known mathematical facts. Similarly, if I tell somebody that I feel a certain way and they fail to understand what I mean, the only way I can help them to understand is to guide them using concepts that I think they already have some grasp of, such as by reference to similar feelings that I think they may have had.

A clearer threat to a definition of PC states that makes essential reference to our capacity to communicate comes from epiphenomenalism.

[1.2] Epiphenomenalism

Epiphenomenalism is the perspective that mental states occur alongside or as a result of brain processes without being causally relevant to how those processes occur. It was famously defended in the work of Huxley:

> The consciousness of brutes would appear to be related to the mechanism of their body simply as a collateral product of its working, and to be as completely without any power of modifying that working as the steam-whistle which accompanies the work of a locomotive engine is without influence upon its

machinery. Their volition, if they have any, is an emotion indicative of physical changes, not a cause of such changes. (Huxley 1874: 240)

Just as a steam whistle has some physical effect but simply has no useful effect on the machinery responsible for the motion of a train, we may believe that mental states are physical effects themselves but that have no important effect on our behaviour or physiological activity. Yet such a view would not be true epiphenomenalism; it would rather be an example of conscious inessentialism:

Conscious inessentialism is the view that for any intelligent activity *i,* performed in any cognitive domain *d,* even if *we* do i with conscious accompaniments, i can in principle be done without these conscious accompaniments. (Flanagan 1993: 5)

Conscious inessentialism simply implies that consciousness is not essential for intelligent activity, which an epiphenomenalist would agree with but on the grounds that *no* physical activity is impacted by mental activity. Huxley stated that there is “no proof that any state of consciousness is the cause of change in the motion of the matter of the organism” (Huxley 1874: 244). As such, it might be argued that epiphenomenalism would entail that it is impossible to communicate our conscious states; if they are causally inert, then it seems puzzling that they should, in certain circumstances, produce verbal or behavioural responses.

It is important here to distinguish the epiphenomenalism of Huxley from what I will refer to as “physical epiphenomenalism”, which describes consciousness as being “nothing useful” (Flanagan 1993: 130) as opposed to being causally inert:

James aptly refers to the epiphenomenalist position as the "inert spectator" view of the mind or the "conscious automaton" theory. The steam whistle is a

> physical effect of the work of the steam engine. Furthermore, it has physical effects. It adds moisture to the surrounding air and sound waves escape. But these effects are minor and of no consequence as far as the subsequent states of the steam engine go. (Flanagan 1993: 131)

It is fair to state that Flanagan may have taken the steam whistle analogy too literally when he went on to say that a conscious state of seeing a red square could be the brain going into “this funny oscillatory state that persists for a few seconds” but that “does nothing useful” (ibid.: 130). The steam whistle may have physical effects, but Huxley believed that states of consciousness have no effect whatsoever on the physical state of an organism (Huxley 1874: 244). William James, in describing the conscious automaton theory, described there being two lives of Shakespeare, one in which there is a nervous system producing certain motions with a hand during certain periods of his life, and another “in which every gleam of thought and emotion should find its place,” but that “the mind-history would run alongside of the body-history of each man, and each point in the one would correspond to, but not react upon, a point in the other” (James 1890: 133). Neither Huxley nor James in their outlines of epiphenomenalist positions seemed to be describing a physical effect, not even a useless one. As such, Flanagan’s physical epiphenomenalism is a distinct position with very different metaphysical consequences.

The difference between true epiphenomenalism and physical epiphenomenalism will be clear when we see challenges that they must both be equipped to respond to posed by key thought experiments that support the case for epiphenomenalism, such as the philosopher’s zombie and the knowledge argument. These were two of the arguments used by Chalmers in opposition to physicalism (Chalmers 1996: 94–106), although, as we shall see, the property dualism of Chalmers does not necessarily entail epiphenomenalism; nonetheless we will consider these cases insofar as they do constitute arguments for this position.

[1.2.1] Philosophers' zombies

The philosopher's zombie is a posited being that is a physical duplicate of a conscious being but that lacks consciousness. Such a being is supposed to be conceivable, whether such beings do exist. As Chalmers stated:

> The idea of zombies as I have described them is a strange one. For a start, it is unlikely that zombies are naturally possible. In the real world, it is likely that any replica of me would be conscious. For this reason, it is most natural to imagine unconscious creatures as physically different from conscious ones exhibiting impaired behavior, for example. But the question is not whether it is plausible that zombies could exist in our world, or even whether the idea of a zombie replica is a natural one; the question is whether the notion of a zombie is conceptually coherent. The mere intelligibility of the notion is enough to establish the conclusion. (Chalmers 1996: 96)

Leaving aside the question of how Chalmers is supposed to have established that such zombies are unlikely to exist in our world (since they are impossible to distinguish from non-zombies), we can follow Chalmers' reasoning to the conclusion that the intelligibility of zombies demonstrates that the fact that conscious experience exists is not entailed by facts about physical states (ibid.: 97).

If, then, it is metaphysically possible for zombies to exist, it follows that there is no way for conscious states to cause us to make true claims about them:

> Suppose, for instance, that Otto insists that he (for one) has epiphenomenal qualia. Why does he say this? Not because they have some effect on him,

> somehow guiding him or alerting him as he makes his avowals. By the very definition of epiphenomena (in the philosophical sense), Otto's heartfelt avowals that he has epiphenomena could not be evidence for himself or anyone else that he does have them, since he would be saying exactly the same thing even if he didn't have them. (Dennett 1991b: 402–403)

Dennett pitted himself against philosophers who claimed that there would be some distinction between the behaviour of a zombie and a non-zombie. Moody argued that zombie philosophers would not understand the Other Minds problem and would likely not have even considered it until they met us (Moody 1994: 198) and Flanagan and Polger argued that it is implausible that mentalistic vocabulary would evolve among zombies (Flanagan & Polger 1995: 315). However, as Dennett noted, zombies would *ex hypothesi* have the exact same vocabulary including all mentalistic and phenomenal terminology and they would behave exactly as if they did understand the Other Minds problem (Dennett 1995b: 172).

The trouble with this, as Dennett noted, is that it means that our reports about our conscious states are completely independent of our conscious states. Our beliefs could not be what is responsible for our saying what we believe because even if we were zombies without the capacity for beliefs, we would still go on to say the same things (Dennett 1991a: 403). Chalmers responded to this criticism by saying, "So what?" (Chalmers 1996: 198). He stated that, in the case where I am reporting my conscious state, I am stating a true belief I have by virtue of clear evidence of conscious states, but in the zombie case, my twin is reporting something that he does not have evidence for and is thus false.

> If I say my belief is justified by my immediate acquaintance with experience, it is noted that my zombie twin says the same. To this, the answer is again, "So what?" At most this shows that from the *third-person* point of view, my zombie twin and I are identical, so that *you* cannot be certain that I am conscious; but

we knew this all along. But it does nothing to imply that from the *first-person* view I cannot know I am conscious. (Chalmers 1996: 198– 199)

There is a clear argument against epiphenomenalism as defended by Chalmers here that follows from Dennett's point. The focus of Dennett's argument seems to be largely that it seems implausible that beliefs could be so radically separate from reports, and other philosophers have expressed incredulity at the same kind of claims Dennett argued against. Batthyány and Elitzur stated that "One has to read this passage time and again in order to believe what it says: *A philosopher writes a book about qualia, discussing their enigmatic nature in great detail, and then states that he would write exactly the same book had he lacked qualia*!" (Batthyány & Elitzur 2009: 17). There is, however, a more direct argument to be made against these claims, which follows from the fact that there must be instances where non-zombies make *true* claims about conscious states that need to be accounted for and it seems that they cannot be under epiphenomenalism.

My zombie twin can say that he is experiencing pain, but this would be false, whereas when I say so about myself it is true. Chalmers' explanation would be that it is the phenomenal quality of pain that enables the claim to be true in one instance but false in the other, but since I would have stated that pain is present whether or not any phenomenal quality of pain actually was present (i.e. whether or not I was a zombie), the presence of pain could have no bearing on the words coming out of my mouth. For my zombie twin, there is a purely physical story that can account for him stating things such as, "I am in pain," and, "You make me feel very uncomfortable." For me, the exact same physical story must be true since we are physically identical. We thus have a purely physical story for how statements about my conscious states are produced that requires no intervention from our conscious states themselves. Yet this seems problematic; if a certain fruit was placed in the same room as me at random intervals and I was able to report accurately when it was present and when it was not, this would only make sense if the presence of that fruit was able to affect my ability to report its presence,

such as by the fruit affecting my sensory organs by direct contact or indirectly via a medium such as lightwaves or chemicals released into the air. Indeed, we would be forced to posit a story such as this because otherwise we would have to conclude that my accurate claims about the fruit's presence were purely coincidental. Along the same lines, we need to be able to account for the accuracy of claims about my conscious states if those states have no impact on my ability to report since without one we resort simply to coincidence. While being in pain may suffice as evidence for me that I am in pain, my capacity to produce a report about it would be totally separate from my knowledge about my mental states.

One might seek to avoid any appeal to coincidence by stating that the content of a zombie's belief is distinct to the content of a non-zombie's belief, as Moody (1994) and Chalmers (1996: 174–175) entertained, and so if a zombie were to say that she was in pain, given that she lacks phenomenal qualities of pain she must be referring to some other non-phenomenal state, such as a particular functional or dispositional state. There might even be some zombie equivalent of phenomenal concepts; they might use a concept in place of "phenomenal red" for which they use the same name but instead denotes whatever dispositional state they happen to be in at a particular time. To make this point the strongest it might be, let's assume that all phenomenal terms can have alternative zombie meanings such that a zombie could join in apparently meaningful discourse about consciousness with non-zombies; there would be a perfectly coherent story we could tell for zombies that would be different to the story we could tell for non-zombies that would describe why they make the claims they do.

Still this would not suffice. If I wish to use the plausibility of zombies to claim that epiphenomenalism is true, I still must be able to account for how, in this world, I am able to truly state that I am in a particular conscious state. The explanation for why my zombie twin produces his statement about his own state (that he describes as conscious) would also be true of my own statement; the exact physical circumstances that produce his claims produce mine. So if there is a full account for how my zombie twin produces his statement that makes

no reference to his consciousness, then there is similarly a full account for how I produce my statements that makes no reference to my consciousness. That means that if there is a story I can tell about there being two sets of facts, one set of physical facts and another set of phenomenal facts, this story could only accurately reflect the truth by coincidence, since they cannot both independently affect my capacity to say words. As such, to express facts about epiphenomenal qualia would require that we do so purely by chance. We cannot accept then that it is a genuine metaphysical possibility that zombies exist or we lose any legitimacy to our ability to make true claims about consciousness, which is surely a precondition for stating that any stance on consciousness, including epiphenomenalism, can be known to be true.

Under physical epiphenomenalism, we do not have this problem because we would have to presuppose that zombies do not exist anyway; if conscious states accompany some particular set of physical states that are useless in some particular functional sense, they are still physical states and thus the prospect of plausibly having a physically identical situation where those conscious states are absent cannot be accepted. If conscious states are physical states that are in some sense inessential to our physiological activity and behaviour, then provided those physical states can still cause us to report our mental states, we would not have to rely on coincidence to explain how we are able to make true claims about consciousness. We could simply state that we are able to report states of consciousness accurately by virtue of the (otherwise useless) physical states that are present wherever there is conscious activity.

[1.2.2] The knowledge argument

The knowledge argument against physicalism was introduced by Jackson:

> Mary is a brilliant scientist who is, for whatever reason, forced to investigate the world from a black and white room via a black and white television monitor.

> She specialises in the neurophysiology of vision and acquires, let us suppose, all the physical information there is to obtain about what goes on when we see ripe tomatoes, or the sky, and use terms like 'red', 'blue', and so on… What will happen when Mary is released from her black and white room or is given a colour television monitor? Will she learn anything or not? It seems just obvious that she will learn something about the world and our visual experience of it. But then it is inescapable that her previous knowledge was incomplete. But she had all the physical information. Ergo there is more to have than that, and Physicalism is false. (Jackson 1982: 130)

The conclusion that Jackson sought in this paper is that qualia are epiphenomenal. He proposed that qualia are "left out of the physicalist story" (ibid.: 30). Some (i.e. Churchland 1990; Dennett 1991a: 398–401) argued against the plausibility of the premises of Jackson's argument, but I will show that this argument cannot demonstrate what Jackson intended simply because, if they were epiphenomenal, Mary would not be able to report qualia at all. If conscious experience was epiphenomenal with no causal descendents, the fact that subjects can make accurate descriptions of their conscious experiences would be inexplicable. If indeed Mary were to become acquainted with epiphenomenal qualia by seeing redness, then the physical state of her mouth moving to form the words, "Wow, so this is what red looks like!" must be causally unaffected by the presence of the phenomenal quality of redness, and hence any facts that she then expresses about her qualia – "Redness is not a physical feature of the world… it is epiphenomenal!" – could only be true entirely by luck. It must simply be a coincidence that the physical state of Mary's brain caused her to produce a set of words that match what is true of qualia, which, by being epiphenomenal, can have no have no effect on the vibrations produced in the air from her mouth.

The trouble with Jackson's position is not that physicalism is necessarily true. It is that epiphenomenalism tries to strike an uneasy balance between accepting the causal closure of

the physical world and arguing that we can know about something existing outside of that causal structure. To admit that there is a causal story that defines the physical world and that there are additionally conscious states is one thing, but to admit that there is such a causal story and that conscious states cannot fit into it makes it baffling that we should be able to make true claims about our conscious states.

To argue that conscious states lie outside our ordinary conception of the physical world then would require some major revision to the physicalist story, and we may look to options such as refuting the idea of causation to blunt attacks against epiphenomenalism. This option may work as a defence of property dualism but should be entirely divorced from efforts to defend epiphenomenalism. We could accept the argument by Chalmers to the effect that, if we take Hume's concerns about causation seriously, we have no reason to suppose that non-physical events cannot be followed by physical events. Hume stated that our perceiving a "cause" and "effect" would either involve us perceiving one event (the cause) happening simultaneously with another (the effect), in which case things would not change over time – "The consequence of this wou'd be no less than the destruction of that succession of causes, which we observe in the world; and indeed, the utter annihilation of time" (Hume 1748: 1.3.2.7) – or involves us perceiving one event happening after another event, in which case all we see is one thing happening and then another, such as one billiard ball moving toward another, being in direct contact with it, and then the other billiard ball moving away). This being so we never see whatever it is that ensures that the second event should follow the first – we never perceive the "necessary connexion" (ibid.: E7.6–E7.8) between the two. What we actually perceive is their "constant conjunction" (ibid.: E7.28); we see that one has always followed the other (although it should be noted that the claim that Hume did indeed hold such a "regularity theory of causation" has been doubted by some, such as Strawson (2014)). Chalmers argued, "If there are indeed [necessary] connections, they are entirely mysterious in both the physical and psychophysical cases, so the latter poses no special problem here" (Chalmers 1996: 170).

Such a move is a way of denying that property dualism entails epiphenomenalism rather than a defence of epiphenomenalism. If we take Hume's concerns seriously, we might state that all observed instances of one event following another are not observations of events necessitating one another, but we would still presumably recognise that there are sets of rules that observed instances of events seem to have followed. Every time I have dropped a glass from a certain height onto solid ground it has shattered. If there were no regularities in nature, this could only be explained as a series of increasingly unlikely coincidences. As such, it is still the case that I wish to state that somehow the glass shattering was determined, at least partially, by my dropping it from a certain height onto solid ground. This may not be a causal law, there may be no necessary connection between the two, and I may not even know that the next time I drop a glass onto solid ground from the same height I will end up with the same result, but due to whatever (possibly contingent) set of circumstances held, I am still able to truthfully state that the dropping of the glass onto the ground must have been a determining factor in its shattering all those times.

Epiphenomenalism, on the other hand, would posit that my stating, "I just saw red," was not determined at all by the presence of the conscious state of perceiving red. This, then, requires that the facts expressed in such statements can only have their factual content by coincidence since the actual presence of the conscious state in no way determined that the statement was made. If we argued instead that the statement was determined by the presence of the conscious state of perceiving red then this position would not be epiphenomenalism; it would just be the same as arguing that the presence of consciousness can result in physical activity, albeit with a modified description of what constitutes resulting in such activity such that causation is not mentioned. Either way, we cannot profess to use Hume's concerns about causation as a way of avoiding the difficulties of epiphenomenalism.

Chalmers used another method to avoid the conclusion that consciousness must stand in a causal relationship to our capacity to report our conscious states, which was to oppose the causal theory of reference espoused by Kripke (i.e. Kripke 1972) and argue that our ability

to refer to things does not depend upon us standing in a causal relationship to them (Chalmers 1996: 201). Chalmers stated that we can, for instance, refer to "the largest star in the universe" without being in a causal relationship with that star; all that matters is that "our concepts have *intensions* that the entity might satisfy" (ibid.: 201), with an "intension" being "a function specifying how the concept applies to different situations" (ibid.: 55). As such, he argued that we can refer to conscious experiences in our discourse without enjoying a causal relationship with them; there just must be criteria that the largest star in the universe would have to meet for us to be able to refer to it as such.

Even if we were to adopt this Russellian account of reference, there is a clear difference between claims I might make about the largest star in the universe and my own conscious states. If I am an expert in astronomy I might know, based on models we have of the universe, the rough chemical composition of the largest star in the universe, and, knowing how large a star could possibly be before it would collapse in on itself or explode, the rough size of the star, all without having any causal relationship to the star. Nonetheless, it should be quite clear that I am referring to any possible number of objects; while there is one star that presumably must fit the label "the largest star in the universe" now, that star may explode immediately after I have mentioned it and hence by the time I use the same label I will be referring to a completely different object without knowing this fact. I could refer to the largest star in the universe at time *t*, but my reference to the star would have seemed the same to me whether the star was one hundred million light years away or twenty billion. The absence of causation between me and the object I am referring to means that, for me to be able to make true claims, my reference must be necessarily vague and able to apply to a vast number of different instances.

On the contrary, the claims I make about my conscious states are claims about states I am in at particular times; they pick out specific existences. I can claim that the pain I am feeling today is not as bad as it was yesterday but is much worse than it was two years ago when I broke my arm. This is not something that would be true for any given pain, it is specific to the pain I am in now. If I can refer to the biggest star in the universe with no causal

connection then my claims must be vague enough that they could be held by any existing star worthy of the title, whereas my claims about my pain can be specific enough that I can make claims about instances of them happening at particular times, and the relative severity of that instance compared to other instances. I can know how large (at a minimum) the biggest star in the universe must be based only on inferences made using my observations of previous stars, but I cannot know how severe a particular pain is just based on having experienced numerous other pains.

With regard to other entities that we can bear no causal relationship to, such as fictional entities, Kripke accounted for these by arguing that the entity is something we pretend to refer to (Kripke 2011: 59). It can be accounted for in terms of a causal theory of reference by stating that since the entity is invented, we simply learn the pretend referring term from other people; fictional entities are not "shadowy possible people" but are rather entities existing within the context of a certain work of fiction (ibid.: 63). Such a theory of reference is clearly not appropriate for our references to consciousness if we believe that consciousness both exists and has no causal effect on our bodies, since it would apply only if consciousness were non-existent. However, even if we reject such an account with regard to abstract and fictional entities and assert instead that conscious states are abstract things we bear no causal relationship to, this still does not give us grounds to assume that they can be referred to non-causally. Even if numbers are abstract entities that I bear no causal relationship to and even though I can still make specific claims regarding particular numbers, I have some means of making such claims *a priori*. On the contrary, I cannot make claims about conscious states *a priori*; I must experience a conscious state to be capable of making true claims about it. It is not *a priori* knowable which conscious state you will be in at a future time any more than it is *a priori* knowable what any specific state of the physical world will be at a future time. I can also make true claims about fictional entities that are not *a priori*, such as by stating that Sherlock Holmes lived on Baker Street. Yet, if I am able to refer to such entities despite bearing no causal relationship to them, it is due to the fact that such entities do not exist; facts about

Sherlock Holmes are exhausted by what was written by the author of the stories and what logically follows from that which was written (i.e. it follows that Sherlock Holmes lived on a street, and that he lived in a place, and that he lived). Since Sherlock is only fictional, no further facts are required to account for what he did or did not do, whereas the same cannot be said about conscious states. If conscious states are exhausted by what we sincerely say about them then, regardless of whether any particular qualitative properties are present, we could report our conscious state at time X by either stating, “I am seeing a red square,” or, “I am seeing a green triangle,” with the same truth value. Thus, the epiphenomenalist cannot take this route otherwise she is no longer an epiphenomenalist but an illusionist.

Suggesting that we make true claims about conscious states without having any causal relation to those states is incompatible with the idea that we can accurately report the conscious states we are in at particular moments. Whatever the metaphysical nature of conscious states, whether they are composed of qualitative properties or not, all the positions mentioned agree that when you say, “I feel warm,” it is because you are in the conscious state of feeling warmth. That capacity to produce reports about specific instances of conscious experience cannot be a vague reference to an abstract or distant property otherwise the accuracy of such statements is inexplicable and we are, once again, forced to assume that they can only be made by coincidence. If, then, Mary can express her surprise at the colour of roses then it is spectacularly unlikely that this is because there are epiphenomenal red qualia.

[1.2.3] Knowledge by acquaintance

One response to the knowledge argument is to state that the kind of knowledge Mary acquires when she perceives a red rose is distinct to the kind of knowledge Mary acquires when she learns about red roses by seeing black and white images and words, or by hearing

descriptions. The distinction is drawn between knowledge by acquaintance and knowledge by description, as introduced by Russell (1910–1911). Knowledge by acquaintance requires a subject to be in contact with phenomena such that this knowledge is a direct result of that contact, as opposed to being inferred as a result of, for example, interpreting what images and descriptions represent. So, when Mary encounters redness she gains *noninferential* knowledge from her acquaintance with this phenomenon. This is a foundationalist perspective, taking as it does that all knowledge is supported by a form of noninferential knowledge.

Knowledge by acquaintance has been related to Mary's case by several authors. Chalmers has stated that Mary gains phenomenal knowledge of a kind that "is prima facie justified in virtue of the subject's acquaintance with that quality" (Chalmers 2010: 286). Bigelow and Pargetter argued, from a physicalist perspective, that knowledge states can change in multiple ways, one of which is at "the level of acquaintance relations" and Mary is a case of somebody whose knowledge state is structurally the same, as it relates to the same part of the world, but where there "clearly is a change in what she knows" (Bigelow & Pargetter 1990: 139). Conee, also from a physicalist perspective, has argued that "knowledge by acquaintance of an experience requires only a maximally direct cognitive relation to an experience" (Conee 1994: 136) and thus that no new information is learned (ibid.: 148).

We can interpret knowledge by acquaintance under either a physicalist or dualist light, as Feigl observed in contrasting monist and dualist interpretations (Feigl 1958: 69–70). The difference between the two interpretations hinges on what it is they conclude we are acquainted with.

A dualist theory, involving a non-causal relationship between us and non-physical states is straightforwardly another example of epiphenomenalism. If knowledge is gained through direct acquaintance with states that are distinct from those with causal efficacy then this knowledge would have no bearing on the words that come out of Mary's mouth when she sees a rose for the first time. To place acquaintance outside of all causal interaction entails that knowledge of these states would be gained without producing any observable physical

effects. As such, any claims made about such non-physical states could not be made as a result of the presence of such states. This brings us back to the problem with epiphenomenalism already discussed; the truth of any claims about conscious states wold be completely independent of the process by which the claim itself was produced. The claims could thus only be true by coincidence.

If we adopt a physicalist interpretation of acquaintance, there would be no reason to suspect the effects of acquaintance not to be observable. Feigl, for instance, stated that knowledge by acquaintance is direct in the sense that "it seems utterly inappropriate to ask someone what his evidence is for asserting that he, e.g., feels at the moment elated, depressed, anxious, dizzy, hot, cold, and so on through the various modalities and qualities" (ibid.: 37). As such, if we are physicalists we would simply need to state that we know ourselves to feel dizzy because we are in whichever physical state involves us feeling dizzy. This need not be evidence of the infallible truth of any particular proposition; we would, to accept this doctrine, need only accept that it is as a result of something we are non-inferentially aware of that we have come to describe ourselves as feeling dizzy and that this non-inferential awareness has occurred because we are in some physical (e.g. physiological) state. Acquaintance, if we are physicalists, could only be either the presence of a physical state or a relation between physical states. In this case, we lose any reason to suspect that this particular set of circumstances should not be observable. Indeed, it could be that Mary simply became acquainted with something new because she had not been in the particular physiological state that results from seeing a red rose before.

Of particular relevance to the current discussion, if acquaintance involves being a particular physical state this should produce physical effects, which should be detectable just like any other physical effect. If we hold this position, then, we should expect this overall physical state to produce effects by virtue of which it can be empirically identified.

The argument from knowledge by acquaintance thus does not give us any reason to suspect that there are conscious states which do not produce observable effects, since it either

supposes epiphenomenalism to be true, in which case I refer the reader back to the the counter-arguments against this position that have already been given, or else it supposes that acquaintance need not entail any non-observable effects by virtue of which these states could be empirically identified.

The purpose of arguing against epiphenomenalism is to justify my description of PC states as being a sensible starting point for the study of consciousness. It has been shown that it is implausible to deny that in PC states consciousness must be causally involved in the production of the behaviours that communicate our conscious states. There has been no refutation of physical epiphenomenalism, since this posits only that conscious states are irrelevant to the functions that produce them, not that they have no effect on our ability to describe them whatsoever. For instance, one could be in a certain neurophysiological state and have an accompanying conscious experience of redness, and although the redness may only be associated with some neurophysiological property not essential for the function of processing information about certain surface textures or wavelengths of light, all that is required to overcome these objections to epiphenomenalism is that it at least serves a function in the role of producing our descriptions of conscious states. The “funny oscillatory state” that Flanagan offhandedly posited as the ontological basis of conscious experience could well fit our definition of conscious states and avoid the problems of epiphenomenalism provided it is plausible that such a state could play a role in producing true claims about conscious experience, and we have not yet been given any reason to doubt that this is possible.

[1.3] Empirically identifying PC states

It follows from the above considerations that PC states should be possible to identify empirically. If conscious states can be causally implicated in a story about a subject’s

production of certain behaviours or a subject saying certain words such as, “I see red,” then this is a story that we should be able to tell through empirical observations of the causes of those activities.

The philosophical problems of consciousness do not arise from a rejection of the claim that we can explain what makes us produce the words and behaviour we do. That is, there is no *a priori* objection to the ability of fields such as neuroscience, biochemistry and cognitive science to describe how even intelligent behaviour can be produced by a biological system such as a human body and indeed our greatest hope of such understanding surely lies in these enterprises. It can be supposed that once a greater understanding of the structure of the human brain and its capacity to process information in its huge and complex variety of ways has been reached, it would be possible in principle to model and predict behaviour in a huge number of situations such that there would be nothing mysterious about the fact that you say, “I saw something just like this when I was on holiday last year,” when you are in the presence of certain stimuli.

Nonetheless, an adequate explanation of what makes those statements true would not adhere to the same criteria as an adequate explanation of what made the words come out of your mouth at that moment. An explanation of why you made a certain statement at a certain time need only include a description of your neurophysiological state at that time and an explanation of what caused that state to change in such a fashion as to produce those words. Yet, for those words to reflect something true, other criteria need to be met; something that happened at the time when you were on holiday last year is also determinate of the truth of the statement. There is thus, in many situations, a distinction between the proximate cause of your statement and the truth-maker of that statement.

Even so, there must be some causal story that we can tell about how it is that you have come to be able to truthfully state that you saw that same thing last year, presumably involving there being an object bearing a certain spatial relation to your body at that time and then your brain being structurally modified in such a way as to be able to produce statements about it at

least up to a year later. There is, again, no mystery to this other than the (presumably) scientifically soluble mystery regarding the mechanisms that allow such a thing to happen.

Whether or not all true statements have both a proximate cause and a causal story that can account for the truth-maker of that statement, and even if statements about certain fictional or abstract entities, such as numbers, do not have both proximate causes and truthmakers, we should expect there to be both when it comes to unambiguously spatiotemporally located instances. An individual making a certain claim such as saying, “I am in pain,” is an example of such.

There should, then, be nothing different about PC states, given that statements about conscious experience describe specific events and temporally (if not necessarily spatiotemporally) located states. If I state truthfully that I am having a certain experience there must be some proximate cause for the production of this statement. For reasons already discussed in the rejection of epiphenomenalism, there must also be a causal story that can be told about how conscious states have resulted in such a true statement. The same applies to statements about past experiences; there must be both a proximate cause for the production of my statement and an overall causal story that can describe how conscious states factored into the chain of physical events that eventually culminated in my producing a true report about them.

PC states, then, should not cause a major scientific problem. Even if they are nonphysical, being metaphysically distinct from physical entities in some important sense, they would still have to be causally efficacious and thus would still be detectable through empirical means. If I can observe an event that is caused by a previous one, then it is presumably *in principle* possible for me to have observed the whole chain of events leading to any observable event, with perhaps the exception of unknown cases such as just before the Big Bang or inside a black hole. To be able to detect both proximate causes and truth-makers of our claims about consciousness is the sort of thing that scientists are currently trying to achieve and I will go into more detail about their approach in the next section. An example of such a method of

study would be to identify the neurophysiological states present alongside paradigmatic states of consciousness as these neurophysiological states must presumably be important in an account of our capacity to produce true claims about conscious states.

There are at least two possible directions one could take regarding the results of such a method of study. One could take it that conscious states can be reduced to certain physical states, such as by arguing that there are reasons related to parsimony of explanation that give us reason to prefer reducing consciousness to a physical state (Feigl 1958: 15). In such a case the form of a scientific study is easy to discern; it would entail identifying the physical state responsible for our ability to make true claims about conscious states. Alternatively, we could take it that our knowledge of our conscious states comes from data that is separate from that gathered by third-person science (Chalmers 2003: 1111) and that, since one form of data cannot be reduced to the other (ibid.: 1112), the best that science can do is to draw up correlations between these two forms of data. Chalmers has stated that, "A science of consciousness will not reduce first-person data to third-person data, but it will articulate the systematic connections between them" (ibid.: 1113). This provides justification for the project to identify the neural correlates of consciousness (ibid.: 1115), which is the most prominent scientific investigation into the physical basis of consciousness to have ever been undertaken and has been prolific in generating research on this. The general scientific approach of which the neural correlates is an example will be considered in some detail in the next section.

For now, it will suffice to say that we do not need to distinguish between property dualism and the identity theory for the purposes of investigating the scientific means of studying consciousness. This is in line with the views of prominent scientists, some of whom have defended a form of the identity theory (albeit without calling it such) (e.g. Crick 1994: 3) and some of whom have defended property dualism (e.g. Koch 2012: 152) without this resulting in any clear incompatibilities between their approaches. It is also unnecessary to distinguish between the two in terms of the research because identities have to be made *a posteriori* given that they rely on finding correlations anyway, as Feigl described when he stated that

identification of phenomenal properties to brain states is “restricted to those elements, properties, or relations in the neural processes which (in dualistic parlance) are the "correlates" of the raw feels” (Feigl 1958: 90) and so the research goals and methods are at least partially the same regardless of the position. Most importantly though, it is not the distinction between these two positions that is of interest here because neither perspective is equipped to deal with the problem that arises for consciousness science.

Nonetheless, it can still be maintained that it must be possible to detect conscious states or their physical correlates; since statements such as, “I taste sweetness,” if true, describe something happening at a particular moment and as such we must be able to determine what it is that has produced our capacity to make such statements. This applies to both the proximate cause and truth-maker of the statement; if my conscious state is one of recalling an experience I had a year ago then my state a year ago must factor into a causal explanation for how the claim about my present conscious state can be produced. It must presumably be possible in principle to have measured the events occurring both in and around me over the last year and so find in the causal chain the states by virtue of which such statements can be true. Denying this cannot be done without some significant revision to our understanding of physical measurement or consciousness.

PC states should thus be possible to identify through standard scientific means. There are two terms that will need clarified here: “identify” and “standard scientific means.”

The reason I use the word “identify” to describe what must be possible to do empirically is that, if I were instead to state that it must be possible for us to observe PC states, what exactly this consists of will depend upon the metaphysical predilections of the reader. For the reductive physicalist, observing a PC state would simply involve observing a brain state, whereas for a property dualist this may not suffice. Regardless, both perspectives must also be capable of determining that a particular conscious state is present alongside particular observable states; it is not simply the case that the reductive physicalist should be able to observe a conscious state by observing a brain state, she must also be able to recognise that

the brain state she is observing *is* the conscious state she is attempting to study. A statement such as, "I see red," must be true by virtue of the presence of the conscious state; if the empirically observable circumstances within which that conscious state is present cannot be ascertained then the reductive physicalist cannot determine its causal effects. Even if conscious states are brain states, identifications between the two must be made *a posteriori*. The word "identify" thus describes what is required for a metaphysical position to recognise the presence of a conscious state in order to determine what its causal role is whether a physical state is caused by a conscious state or is identical to it. It is also suited to describe what a property dualist would suppose that a science of consciousness could do; it would suppose that a science of consciousness could identify a PC state by the presence of a neurophysiological correlate, for instance. As such, the term "identify" does not presuppose a particular metaphysical position on the physicality of consciousness.

The second term that will need clarification is "standard scientific means." I will give further explanation of this term in [2.1].

[1.4] Summary

We have so far highlighted the specific kind of conscious state that will be the object of investigation throughout this thesis, which are paradigmatic states of consciousness, given that these should admit the possibility of identification through empirical observation. We have explored the possibility that such states are non-characteristic of consciousness on the grounds that conscious states could be causally inert, and rejected it as resulting in true claims about consciousness being possible to make only by chance. This has vindicated our view that PC states should be our focus if we wish to study consciousness empirically, and it is to this study that we shall now turn our attention.

2

Attempting to Study Consciousness Scientifically

[2.1] Standard scientific means

To be able to claim that something must be possible to detect through standard scientific means requires some idea of what must be required of a practice for it to be considered a standard scientific one. This may seem to presuppose not only that it is possible to determine what criteria a study must have in order to be scientific, and thus solve the problem of demarcation (Popper 1935: 11), which anybody who endorses the kind of anti-method picture of science painted by Feyerabend (1975) will heavily dispute, but also that I will be able to do so in order to support my statement, which will contradict the institutionalism of Ladyman, Ross and their collaborators (2007), which states that something only counts as scientific knowledge if it is considered such by scientific institutions. However, the claim that conscious states must be possible to detect through scientific means does not need to contradict Feyerabend's position and there is good reason to reject institutionalism.

[2.1.1] Anarchism

Feyerabend's position in Against Method is that "*the events, procedures and results that constitute the sciences have no common structure*" (Feyerabend 1975: 1). Throughout his work, he attacked the notion that science relies upon facts, rational justification or empirical observation, citing examples throughout the history of science where each of these

paradigmatic descriptions of the scientific process have been mistakenly or intentionally ignored.

Much of the attack upon Feyerabend's work has been based around the implication of relativism, since he stated that there is no methodology that can be ruled out before science can be conducted, yet in a postscript to his book added later in his life he described his stance as being that there are a variety of "often contradictory" methods that scientists adopt but "not every approach succeeds" (Feyerabend 1993: 270).

As such, his view is not an example of relativism but is better described as "anarchist pluralism," (Wartofsky 1991: 28) as he takes it not that any approach must be measured by its own set of standards, but that there is no singular approach that should succeed in portraying reality. Nonetheless, if Feyerabend was correct that only certain practices can be successful, then this must be according to some standard of success. Although there may be no specifiable scientific methodology, if there is no standard of success in science then this allows the distinction between science and non-science to be so loose as to prevent the word "science" from referring to anything at all. If one allows for no standard of scientific success, it would be possible to argue not only that astrology, parapsychology and tarot reading are examples of scientific study but also that bouncing a ball against a wall or brushing one's teeth count as scientific study, given that these practices could be considered successful within *some* standard of success (such as relieving boredom or avoiding cavities). Indeed, if one is to argue that there is no standard of scientific success and no scientific method, it would seem tactically to make more sense to simply be an eliminativist about science; if something has no defining features whatsoever then this implies that references to that thing fail to capture anything at all.

As such, even if there is no method we must be able to distinguish between scientific success and non-scientific success, but this task again requires us to demarcate science from non-science. Even without performing this difficult task, I think it is possible to specify one standard of success that is necessary for something to be a scientific practice, even if it is not

sufficient for something to be a scientific practice. This is all that will be required for the work to be done throughout this thesis because it is possible to demonstrate that attempts to study consciousness scientifically fail to meet even this standard, which demonstrates that such attempts are failing.

This standard of success for a scientific practice is that it provides a means of either accepting or rejecting propositions based on what is found in observation. That is, it must be *in principle* possible for an observation to affect our knowledge of the likelihood of a given proposition being either true or false. It is not the sole standard of scientific success, such that whether a scientific theory is more successful is determined wholly by its ability to can accept or reject more propositions on the basis of observation than its predecessors, since it may well be the case that some theories are rejected in favour of other theories that accept falsified statements as fact:

> what shall we make of the methodological demand that a theory must be judged by experience and must be rejected if it contradicts accepted basic statements? What attitude shall we adopt towards the various theories of confirmation and corroboration, which rest on the assumption that theories can be made to agree with the known facts, and which use the amount of agreement reached as a principle of evaluation? This demand, these theories are now all seen to be quite useless… In practice they are never obeyed by anyone. Methodologists may point to the importance of falsifications – but they blithely use falsified theories, they may sermonize how important it is to consider all the relevant evidence, and never mention those big and drastic facts which show that the theories they admire and accept may be as badly off as the older theories which they reject. (Feyerabend 1975: 50)

Yet the standard outlined above can remain a reasonable standard of scientific success even if, as Feyerabend argued, sometimes old theories are rejected in favour of new theories despite both having the same evidence in their favour. My argument is not that the success of scientific theories is wholly determined by their capacity to describe observable differences, but is rather that one requirement for a science to be successful is for it to have the ability to make distinctions between propositions on the basis of observations. Its ability to do so is *necessary* for its success even if it is not sufficient. Indeed, it may well be the case that new theories are accepted in place of older theories even if new theories accept already falsified claims, as perhaps a new theory will be able to empirically distinguish between propositions that researchers at the time consider to be more important for their contemporary research projects despite the new theory failing to empirically distinguish between many other propositions that were once considered important. It may even be a mistake altogether for a new theory to be accepted, as it may simply fail to empirically distinguish between propositions but be accepted for any of the other reasons Feyerabend suggested. All that matters for the point I wish to make is that, for a theory to be considered a scientific theory at all, it must at least be capable of falsifying or verifying propositions on the basis of observation.

It is also important to mention Feyerabend's argument that what counts as observation depends upon changing views of cognition and thus may be context dependent (ibid.: 51–52). Yet, it can be true that observable verification is a vital part of scientific success if propositions can be rejected or accepted on the basis of observation even if what qualifies as an observation differs in different circumstances. This simply implies that there is no measure of something being an observation that applies across all intellectual contexts and all fields of study, not that the term "observation" is meaningless within fields of study existing at a particular time. It is possible, then, to argue that attempts to study consciousness scientifically fail to allow for propositions to be rejected or accepted on the basis of observation, defined according to its usage within those fields of study.

To unpack the earlier statement about identifying conscious states, we should state this:

> ***Claim A:*** PC states must, in principle, be possible to identify through standard scientific means such that statements about consciousness must be possible to verify or falsify on the basis of observation.

To state that this is the case, then, would not be in opposition to Feyerabend's position, and if this is not in opposition to anarchism then neither is the opposite claim:

> ***Claim B:*** PC states cannot, in principle, be identified through standard scientific means and so statements about consciousness are not possible to verify or falsify on the basis of observation.

I have yet to give a reason for us to take Claim B seriously, but before I do so I will need to show why we should not endorse institutionalism and reject the possibility that such a claim could be made at all.

[2.1.2] Institutionalism

Institutionalism states that science is not defined according to its methodology but can be "demarcated from non-science solely by institutional norms" (Ladyman, Ross et al. 2007: 28). The position is grounded in the notion of the epistemological superiority of science resulting from error filters existing in certain institutional processes:

> Since science just *is* our set of institutional error filters for the job of discovering the objective character of the world – that and no more but also that *and no less* – science respects no domain restrictions and will admit no epistemological rivals (such as natural theology or purely speculative

metaphysics). With respect to anything that is a putative fact about the world, scientific institutional processes are absolutely and exclusively authoritative. (ibid.: 28)

Ladyman, Ross and their collaborators justified this with the accusation that much contemporary metaphysics either ignores or tries to oppose contemporary science. In criticism of Kim's work, they stated that, "As well as ignoring physics, Kim, and much of the metaphysical philosophy of mind of which he is a prominent exponent, ignores most of the interesting questions about the mind that scientists investigate" (ibid.: 18). They pointed to examples of contemporary metaphysical disputes between those maintaining that the world is constituted either of a continuous gunk or divisible into atoms (ibid.: 20), with the former finding no support from anything in contemporary physics whatsoever and the latter relying on a notion of atoms that is well over a century out of date but simply remaining confident that it will deliver atoms of the sort required for their arguments to work (ibid.: 22). They stated that metaphysics relies on intuitions with everyday practical use and that these intuitions cannot "produce systematically worthwhile guidance in either science or metaphysics" (ibid.: 10). They stated further that "no hypothesis that the approximately consensual current scientific picture declares to be beyond our capacity to investigate should be taken seriously" (ibid.: 29).

If this view were correct, this would make it impossible for me to make the claim that science cannot produce testable hypotheses about PC states. Since what determines whether something is a scientific hypothesis, according to the authors, is whether bona fide scientific institutions say it is, it follows that as a philosopher (and not part of a scientific institution) I am in no position to contradict the claims made by scientists whose work is endorsed by such institutions that they are indeed studying consciousness.

If science is to be defined not according to a particular methodology but according to institutional error filters, then all we need to know if we are to identify a hypothesis as scientific is whether or not it "is taken seriously by institutionally bona fide science at *t*" (ibid.: 38). We

need to know, then, what counts as institutionally bona fide science. The authors regarded the Royal Society to be an example of a scientific institution, but Plato's Academy not to be (ibid.: 35), and so there must be some criteria to distinguish one from the other in terms of the legitimacy of their claim to science. However, the authors seemed unwilling to give such criteria, as they wrote, "A remaining locus of ambiguity in our version of verificationism is its appeal to 'institutionally bona fide' science and scientific research funding bodies. In general, we are happy to leave this open to the rational judgements of observers of institutional processes" (ibid.: 36).

Their position can be thus seen as an analysis of existing science and as drawing conclusions from that science rather than attempting to describe what is required for something to count as science. As the authors said in the conclusion of the same book about their defence of verificationism, "Its purpose is to describe limits identified by scientists, not to prescribe such limits" (ibid.: 310). The problem, then, is that we cannot use their criteria to define what constitutes science as opposed to non-science or what constitutes a scientific institution as opposed to a non-scientific institution. If, for instance, I wished to argue that Plato's Academy did actually conduct science or that the Royal Society does not, there is no way to argue against this using any criteria set out in this book; although science as defined by the Royal Society may differ from that as defined by Plato's Academy, this is simply a distinction between the accepted norms of two institutions.

The position expressed in this book can only be defended if we accept that we already know what sort of institutions count as scientific ones, since the book constitutes an analysis of the sort of work produced by certain kinds of institution. For instance, the authors argued that they could make a generalization about institutional practice in the special sciences, which is that "all special sciences stabilize location of the real patterns they aim to track by constructing particular role-fillers to exemplify the cohesion relation" (ibid.: 249). Such a generalization is not something that is *required* of something to be a social science, it is something that they have analysed as a general pattern followed by the social sciences, and

so identifying the social sciences is, for them, epistemically prior to forming a judgement about the operation of the social sciences.

Whatever the value of identifying entrenched institutional practices in the sciences, it cannot be the case that we, or indeed the authors of the book, identify an enterprise as being scientific if it follows such practices; it assumes that certain prominent scientific institutions practicing certain methodologies, such as those adopted in physics, are examples of science and then extrapolates from the work produced by such institutions what constitute the relevant institutional error filters of such a practice. As such, whatever the criteria for the identification of scientific practices such as physics are, they must have been made according to some criteria that could be determined prior to an understanding of institutional error filters. Yet this contradicts claims made by the authors:

> Thus science is, according to us, demarcated from non-science solely by institutional norms: requirements for rigorous peer review before claims may be deposited in 'serious' registers of scientific belief, requirements governing representational rigour with respect to both theoretical claims and accounts of observations and experiments, and so on. (ibid.: 28)

It contradicts this claim because we are required to identify a field such as physics "solely" by institutional norms, which cannot be the case if we only determine what those norms are by looking at institutions that we have identified as scientific in the first place.

When it comes to identifying what scientific institutional features are, we have two options. Either we identify them *a posteriori* by looking at the practices of certain institutions, or we do so *a priori* according to some epistemological justification of what would be required for something to count as an institution of scientific enquiry.

It is clear beyond doubt that the authors strongly opposed the second option. For them, what determines whether something is a scientific hypothesis is whether it is taken seriously

by bona fide scientific institutions (ibid.: 38), not whether or not it meets some specifiable epistemological criteria capable of being met regardless of how any institutions function. On the other hand, identifying them *a posteriori* requires circular justification. If we define scientific practices according to endorsements by scientific institutions, then we cannot identify scientific institutions by their conducting certain scientific practices. So if we cannot look at their practices to identify them, it is difficult to know what we should use as criteria. Indeed, even if we did identify such criteria, such as whether or not an institution can produce results that allow us to make more accurate predictions in future observations, we would again have to identify this either on the basis of some *a priori* epistemological criteria of what constitutes science or on some *a posteriori* practices of a discipline we have somehow determined is a scientific enterprise, in which case we are no closer to understanding how we have determined that it is scientific in the first place.

In short, for institutions to be authoritative over whether or not something is a scientific hypothesis, we must be able to determine what gives those institutions that particular kind of authority, whether this is a certain kind of result or practice, and if we do so then *this* will be the criteria by which we determine if something is science, whether or not that practice is specifically endorsed by a scientific institution. Institutionalism thus does not give us a viable means of demarcating science from non-science. Institutions should rather be considered proxy indicators of proper science (whatever that may be) being practised, in the sense that we may consider a study endorsed by an institution as likely to have been conducted to a certain standard that we have come to expect of that institution, but this does not give us reason to consider the institution to be absolutely authoritative any more than knowing that one of my friends is the most intelligent gives me reason to trust her every utterance to be factual and correct. I base this view on something hinted at by the authors when they suggested that fundability is a proxy indicator for scientific research (ibid.: 34). I agree with this, but the authors again missed the mark when they stated that seeking funding "from nonstandard sources, is, by our lights, precisely the most reliable indicator that their activity

should not be taken seriously by the metaphysician" (ibid.: 36), where a supposed distinction between standard and non-standard sources seems to have been presupposed rather than established on the basis of any specified criteria.

By "standard scientific means", then, I am referring to the capacity to verify or falsify statements according to observation. I do not take this to be what constitutes science, but I take it to be a precondition for something to be science. This means that if a study fails to demonstrate the capacity to verify or falsify statements according to observations it cannot be considered scientific.

[2.2] Recent scientific study of consciousness

Our discussion about the current scientific approach to studying consciousness cannot be divorced from the views and research that actually shape that scientific approach. It is important to base any criticisms about the scientific approach on the structure of existing research and not just on an idealised conception of the conduct of scientific research.

When I use the phrase "current scientific approach to studying consciousness," I mean this to be shorthand for "current scientific approach that purports to be about consciousness." In fact, despite using words like "consciousness" and "qualia," and despite there being a great deal of rhetoric in scientific works about bold new attempts for a scientific approach to begin tackling issues about consciousness, nothing that could legitimately be called a science of consciousness currently exists. Rather, facts are established about cognitive and physiological systems and are then are assumed by extension to be facts about consciousness without any viable empirical methodology established to link these phenomena. This conclusion can only be established following an outline of current and recent scientific practice claiming to be about consciousness.

The two perspectives I will outline will be Baars' global workspace model of consciousness and the theory of thalamocortical interactions outlined by Crick. They have broth proven to be highly influential and, although their conclusions are not universally accepted by the scientific community in the form they were originally presented, it is the methodology that is our focus here. They also provide an interesting contrast between Baars' focus on cognitive science and Crick's on neuroscience.

[2.2.1] Bernard Baars and the global workspace theory

As neuroscientist Daniel Bor wrote:

> Over the past twenty years, the most prevalent, popular psychological theory of consciousness has been the "global workspace theory" proposed by Bernard Baars… Baars' most bold and interesting claim is that, more or less, consciousness boils down to the information sitting right now in our working memory (Bor 2012: 135-136)

Baars' work used a method he referred to as "Contrastive Analysis," which shall be shown to be the essential method used throughout consciousness science research. He outlined this method as follows:

> Now, we can compare a reliably reported conscious image of this morning's breakfast to the memory of breakfast *before* it became conscious; a conscious stream of speech may be compared to the same stream when it is not attended (there is considerable evidence that unattended speech is nevertheless processed up to a point); we can also compare a conscious interpretation of

> an ambiguous word to the same word when the interpretation is not conscious, because the alternative meaning is being accessed (again there is evidence that unconscious word meanings are still briefly processed); or a barely subliminal stimulus may be compared to one presented above threshold; or a habitual unconscious action may be compared to the same act before it fades from consciousness... In all these examples we know that *both* the conscious *and* the unconscious cases involve a mental representation of a very similar stimulus that is apparently processed in a comparable way. Thus, each pair of cases creates a controlled experiment with consciousness as the dependent variable. (Baars 1988: xvii)

Two states are compared, where one is conscious and the other is not; the goal is to identify which characteristics conscious states have when compared to non-conscious states. How we process information consciously is then contrasted with how we process it unconsciously.

Contrastive analysis seems on the face of it to be the most sensible approach to an empirical study of the physiological or cognitive basis of consciousness. As neuroscientist Ramachandran stated a decade and a half later, “We can make a list of all brain events that reach consciousness and a list of those brain events that don't. We can then compare the two lists and ask whether there is a common denominator in each list that distinguishes it from the other” (Ramachandran 2004: 30). After all, if certain states are accompanied by consciousness and others are not, then the best way to determine what features those accompanied by consciousness have in contrast to those that do not is to find examples of both and identify the difference between the two. This is what contrastive analysis aimed to do, and Baars attempted to avoid (for now) ambiguous borderline cases where we are uncertain about the presence of consciousness (Baars 1988: xvii–xviii). Baars adopted this approach in an effort to sidestep philosophical issues and simply find answers to “empirically

decidable ones" (ibid.: xvi). This means that the conscious states Baars is interested in will be only PC states, those that we can reliably indicate through our behaviour and reports.

Even some philosophers who believe that consciousness lies, in some important sense, outside the scope of scientific investigation have seen this sort of approach as valuable in contributing to our understanding of consciousness (Chalmers 1998). While some questions may have to remain unanswered by scientific methodology, such as the "hard problem" (Chalmers 1995), these correlates can still be drawn if we utilise verbal reports as evidence of the presence of conscious states (Chalmers 1998: 221). Chalmers wrote that we must rely on certain "*preexperimental bridging principles*" (ibid.: 220) to conduct such a study, pointing to experiments that aim to draw conclusions about the consciousness of monkeys by seeing if they press a bar in an appropriate way in response to a stimulus. In such studies, experimenters rely on the assumption that verbal reports are simply one example of an arbitrary response necessary to demonstrate the presence of consciousness, with the underlying principle that should be accepted being that it is "when information is *directly available for global control* in a cognitive system, then it is conscious" (ibid.: 222). Chalmers argued that we can rationally reconstruct "neural correlates of consciousness" research so that we accept, as a bridging principle, that consciousness accompanies global availability, and then we do empirical work to discover which neural processes are present where global availability is present, at which point we conclude that consciousness is present where these neural processes are (ibid.: 223). This rational reconstruction seems to be a vindication of Baars' suggestion of 'bootstrapping'. This consists of using reports on consciousness as the primary source of evidence up until correlates have been drawn between the states that have been identified as taking place alongside a subject's reports on their consciousness (for example, neural states), in which case these newly identified states become acceptable evidence for the presence of conscious states even if reports are absent (Baars 1988: 17). Much contemporary consciousness science research can be shown to have used this bootstrapping method and we will encounter some examples throughout this chapter.

To arrive at his model, Baars contrasted various conscious and unconscious processes, such as memorised stimuli that are rehearsed and unrehearsed (ibid.: 22), attended and unattended perceptual streams (ibid.: 20) and deliberate speech and abstract rules of syntax (ibid.: 12). These are examples of where consciousness on the whole is present but the processes involved in one group of them do not seem to enter into conscious awareness whereas the processes involved in the other do seem to. Baars went on to point out that those cognitive processes present in the cases of the conscious states but not present in the cases of the unconscious states have crucial distinctions in terms of their functionality, determining that the unconscious processes mentioned are computationally highly efficient, isolated and autonomous, and operate in parallel with great capacity as opposed to conscious states which are computationally inefficient, with great ability to relate conscious contents to other contents, and operate serially with limited capacity (ibid.: 75). From such considerations, Baars proposed a global workspace model, drawing on theories of information centralisation, to hypothesise that conscious states are those in which multiple processes across a cognitive system can communicate together (ibid.: 86-89).

[2.2.2] Francis Crick's "astonishing hypothesis"

Crick presented his own views as simply hopeful precursors to a developed neuroscientific study of consciousness and he has been described as somebody who organised new fields such as consciousness to make them "scientifically respectable" (Aicardi 2016: 86) rather than the discoverer of the true correlates of consciousness.

In his 1994 book, Crick first touched on a proposal for identifying potential neural correlates of consciousness when he wrote that there is evidence to show that the only brain activity that reaches consciousness is sustained for some minimum amount of time (Crick 1994: 71-72). He went on to claim that we could not have consciousness without some form

of very short-term memory (ibid.: 238), although he offered no argument and discussed no evidence for this in any section of the book.

Nonetheless, working memory and consciousness have been frequently placed together. Baars and Franklin pointed out that all aspects of working memory are reportable, and that reportability is also "the standard operational index of consciousness" (Baars & Franklin 2003: 166). As such, we should not be surprised to see that the two correlate; indeed, given that reportability is such a prominent indicator of the presence of states of consciousness, it is inevitable that anything that correlates strongly with reportability will also correlate strongly with consciousness.

Crick went on to state that a particular thalamocortical connection is involved in shortterm memory (Crick 1994: 240), and he implicated cortical layer 4 at one side of this connection (ibid.: 251). He discussed multiple pieces of evidence, but he stated his view of an ideal method of measuring consciousness when he wrote:

> The evidence so far is so weak that it is reasonable to ask: Can one study exactly the same neuron when the animal is alert and then again when it is unconscious? It is difficult, for technical reasons, to do this if the animal is made unconscious by an anaesthetic, but it has been done by comparing an alert cat with the same animal in slow-wave sleep (ibid.: 224).

In the same section, he stated that ethical concerns make it difficult to conduct such research on humans (ibid.: 228). It is clear that his ideal method would have been to contrast unconscious and conscious states in humans, but that he considered there to be practical and ethical limitations to this method and so results were extrapolated from the closest studiable estimates of these states in other animals such as cats. This was an explicit endorsement of the same sort of contrastive methodology used by Baars.

He pointed out that the biggest change in the cat's cortical activity between it being awake and asleep is in lower layers (5 and 6) (ibid.: 225), but he went on to rule out cortical layer 5 as a likely candidate when pointing out that there is a major pathway from pyramidal cells in cortical layer 5 that seems to be involved in "unconscious" actions (ibid.: 236-237). There are apparently reverberatory circuits running between cortical layers 4 and 6, and so much activity implicated in cortical layer 6 is also implicated in layer 4 (ibid.: 251). He also stated that when a human subject's thalamus is given a certain amount of stimulation, they are able to report awareness of that stimulation (ibid.: 229). The combined implications from links to memory, the thalamus' role in producing reports of conscious experience, and the role of cortical layers 4 and 6 in wakefulness are some of the things that led Crick to his conclusion. Crick regarded any conclusions that can be derived from the research so far as tentative. He wrote:

> I hope nobody will call it the Crick (or the Crick-Koch) Theory of Consciousness. While writing it down, my mind was constantly assailed by reservations and qualifications. If anybody else produced it, I would unhesitantly condemn it as a house of cards. Touch it, and it collapses. This is because it has been carpentered together, with not enough crucial experimental evidence to support its various parts. Its only virtue is that it may prod scientists and philosophers to think about these problems in neural terms, and so accelerate the experimental attack on consciousness. (ibid.: 252)

Based on the above considerations and other examples of experiments further implicating specific areas of the brain, he concluded:

> Consciousness depends crucially on thalamic connections with the cortex. It exists only if certain cortical areas have reverberatory circuits (involving

> cortical layers 4 and 6) that project strongly enough to produce significant reverberations. (ibid.: 252)

Crick's approach was essentially the same as Baars'. As the experiments mentioned above have shown, the goal was to study neural states when an individual is conscious and unconscious and determine the difference, but there were practical difficulties in doing this. As such animal studies such as those involving the waking and sleeping brains of cats were utilised and neurophysiological correlates of very short-term memory were identified, which is an example of bootstrapping, since a link between short-term memory and consciousness is made and then neural correlates of short-term memory are taken to be plausible neural correlates of consciousness. In addition, other contrastive methods are performed, such as where stimulation experiments show which areas of the brain elicit conscious awareness as opposed to not when stimulated. These were all either direct examples of contrastive analysis or were bootstrapped from previous research where such a contrastive method has been conducted.

[2.3] The task of consciousness science

It is interesting to note that neither Baars' model is purely cognitive nor is Crick's purely neurophysiological; there is a great deal of crossover. Baars' global workspace model was partly motivated by attempts at modelling the brain as a system of distributed processing networks (Velichovsky 2017: 37), and Crick's research utilised findings of cognitive science such as the link between working memory and consciousness. While it may be philosophically useful to consider the two separately, the scientific study need not distinguish between the two so long as it is possible for there to be common content between the two areas with implications regarding each other.

It is also of note that neither Crick nor Baars' stances have persisted unchallenged in the scientific community. Some researchers have stated that the prominence of the thalamus in consciousness has been disputed since there have been many identified instances of a change in conscious state while the thalamus remains inactive (Havlík 2017: 80), but this is still an example of the contrastive analysis approach; observations of the presence of a conscious state are contrasted with observations of the absence of a conscious state and the underlying neurophysiological state of each is investigated. Likewise, Baars' approach has been challenged by researchers claiming that the global workspace model fails to differentiate between various kinds of information processing, but they expanded on this model by proposing a distinction between different kinds of conscious information processing rather than disputing Baars' description of the distinction between conscious and non-conscious processing (Song & Tang 2008: 789–793). Models change and research is updated, but the essential approach remains the same.

The growth of research in this area has been entirely consistent with the contrastive analysis approach. Areas that have received much attention since have included understanding of phenomena such as visual masking, where a subject's awareness of one stimulus is interfered with by the presentation of another visual stimulus (Bachmann & Francis 2014: 1), and binocular rivalry, where two incompatible images are presented, one to each eye, with subjects reporting awareness of only one stimulus at a time (Fairhall, Hamm & Kirk 2008). Both these areas contrast "unconscious" visual processing, that is either masked or in the unattended part of the visual field, with conscious visual processing and thus they both involve the methodology used by Baars.

There are many other examples of processes or phenomena associated with conscious activity that are contrasted with those associated with unconscious or nonconscious activity, such as in research comparing anaesthesia and dreamless sleep to wakefulness (Edelman & Tononi 2000: 2; Bor 2012: 81; Damasio 2010: 159), unconscious activity in the cerebellum to conscious activity in the thalamus (Bor 2012: 159-160), driving without awareness to attending

to a conversation (Ramachandran 2004: 31-32; Bor 2012: 110), physiological activity of somebody in a coma to a waking person (Blakeslee & Ramachandran 2005: 238; Damasio 2010: 235), somebody undergoing an absence seizure to somebody awake but not undergoing an absence seizure (Damasio 2010: 163-164) and cognitively processing habituated stimuli (that have faded from consciousness because you are used to them) compared to stimuli you are aware of (Baars 1994: 190-191). A condition known as blindsight has also received much academic focus, where an individual with damage to a certain part of their visual cortex will report no experience in a part of their visual field and then, when prompted to guess, such as being asked to post an envelope through a diagonal slot in the "blind area" of their visual field, will correctly orient their hand to post the envelope properly (Blakeslee & Ramachandran 2005: 64-65). This is contrasted with normal conscious vision, and has been widely discussed as an example of visual processing in the absence of conscious experience (e.g. Farah 1995: 71; Kentridge, Heywood & Weiskrantz 1999: 1810; Silvanto 2008: 2870; & Cowey 2010: 18), but it gradually became unclear as to whether subjects were reporting having no experience at all or simply having a less detailed or degraded experience (Overgaard 2011). Although the possibility that blindsight was never a clear empirical example of a dissociation between consciousness and visual discrimination behaviour was proposed as far back as 1983 (Campion, Latto & Smith 1983), the attention it received is a clear indicator of the sort of research practice favoured in scientific study of consciousness.

Consciousness science, then, proceeds by identifying empirically observable correlates of PC states. These can only be found by contrasting empirically observable states occurring alongside PC states and those occurring alongside states that are not paradigmatic conscious states. Cortical layers 4 and 6 as opposed to cortical layer 5 were outlined in Crick's work as likely correlates of consciousness because cortical layer 5 is also implicated in states that are not PC states.

The methodology, in its own stated terms, seems perfectly viable. If we wish to identify which empirically observable states correspond to conscious states, we must contrast those

present when PC states are present with those present when they are absent. Our physiological and cognitive states are complex and varied when we are in a PC state, but it seems implausible that all parts of these physiological and cognitive states are implicated in that conscious state. Moreover, if an aspect of our overall physiological state is present when we are in pain, for instance, and then continues to be present when we are no longer in pain, we would not be able to identify this aspect of our physiological state as sufficient for pain, and it may indeed even not be necessary for pain. For a state to be a correlate of a conscious state, it must be present where that conscious state is present but absent where it is absent.

This is why the current research method is not only valid but is also the only prospect for a correlation-based consciousness science approach. If we wish to be able to say that we are in conscious state C because we are in empirically observable (cognitive or physiological) state E, then we must be able to isolate which state it is that is present when we are in that conscious state as opposed to when we are not in that conscious state. If E is present when C is absent then E cannot be sufficient for the presence of C.

Any full picture of an empirically observable correlate of consciousness will describe a state that is present only where consciousness is present and not where it is absent.

Broadly, we can outline the task of consciousness science:

> **Consciousness Science's Task:** To discern which empirically observable states (of which our physiological states are an example) are common to instances of PC states but not to their absence.

If consciousness science were to continue in this direction and this were indeed to suffice as a means of finding correlates of consciousness, it would presumably be capable of determining with great accuracy which PC states correlate with which physiological states. Our first glimpse of ambiguity related to the conclusion that can actually be drawn from this

form of research comes when we attempt to establish by what means scientists claim to be able to identify whether or not a state is a PC state.

[2.4] Recognisable conscious behaviour

The above outline of the scientific approach to the study of consciousness requires observation of the presence of PC states to be possible; we must be able to observe a behavioural, physiological or cognitive state and ascertain that it is accompanied by a PC state. It will pay here to go over why PC states are so important here and why a level of ambiguity arises if we wish to talk about conscious states more broadly.

There is no universally accepted set of criteria by virtue of which a scientist can determine whether or not a state is conscious, but voluntary reportability seems to have played a significant role in the capability of scientists to make such judgements (Baars 1988: 12). That is, a subject is considered to be in a conscious state if they are able to voluntarily report being in such a state. I will thus consider voluntary reportability to be an example of a *recognisable conscious behaviour*. A PC state is one that produces recognisable conscious behaviour:

> **PC States:** Conscious states that an individual can indicate the presence of through behaviour or reports.

Where this criterion breaks down is when applied to certain ambiguous situations and non-human animals. Studies of visual systems involved in consciousness have included research performed on monkeys where the monkeys pressing a bar in response to a stimulus is used to indicate the presence of a conscious state, which Chalmers suggested indicates that the criterion used is actually that information is available for arbitrary response (Chalmers 1998: 222). In such a case, we might take the monkey pressing a bar to be a voluntary report,

but the line between what constitutes such a report and what does not seems certainly more ambiguous if we allow for arbitrary responses. There are other cases where again we have to adapt how we are to define reportability, such as cases of split-brain patients. Many individuals with severe epilepsy have had the corpus callosum separating the two hemispheres of the brain severed to prevent seizures spreading from one hemisphere to the other (Taylor & Regard 2003: 257). Although there are very few noticeable differences in behaviour between somebody who has undergone this treatment and somebody who has not, in experimental conditions where participants are shown a stimulus at the far sides of their visual field, subjects with split-brains will report having no experience of stimuli shown to the far left of their visual field. This is presumably because visual information on the far left is processed by the right hemisphere, but the right hemisphere lacks verbal articulatory ability (ibid.: 258–259) and thus in most cases it is impossible to produce a verbal report about information only accessible by the right hemisphere (Sperry 1968: 725). However, when asked to point to the stimulus presented to them out of a list of different options, the individual will point to the correct stimulus, presumably because the right hemisphere still retains control over some motor capabilities (ibid.: 725). When then asked why they selected this option, subjects will typically confabulate a response, such as by saying, “I just guessed.” (ibid.: 726)

In such a case, there is a sense in which a report is still produced when stimuli are presented to the right hemisphere; the subject can indicate that they perceived the stimulus by pointing but not by speaking. Nonetheless, their verbal responses indicate that in another sense they are incapable of producing a report about their stimulus. One way that this seeming discrepancy could be resolved is to consider both hemispheres to be partially independently reporting the information they are processing, which seems to be what Nobel prize-winning neuropsychologist Sperry suggested:

> Instead of the normally unified single stream of consciousness, these patients behave in many ways as if they have two independent streams of conscious

> awareness, one in each hemisphere, each of which is cut off from and out of contact with the mental experiences of the other (ibid.: 724).

If we were to adopt such an interpretation of split-brain phenomena, and if we were to consider the information available for arbitrary response as indicators of the presence of a conscious state, as Chalmers has suggested, we must decide that even if information is available for one kind of arbitrary response, such as the ability to enable a subject to point accurately at a stimulus, it need not be available for another kind, such as the ability to verbalise which stimulus was seen, in order to be considered evidence of a conscious state.

If, on the contrary, we were not to adopt such an interpretation and were to assume that the hemisphere incapable of providing verbal reports was non-conscious, we would have to rule out certain kinds of availability for arbitrary response, such as the ability to enable a subject to point accurately at a stimulus, as being evidence of consciousness. Either way, it seems difficult to make a judgement based on empirical evidence alone; rather we have to determine whether or not we consider one type of availability for arbitrary response sufficient for consciousness and then apply the results of these considerations to empirically observed cases.

It is worth noting that, like blindsight, there is reason to think that split-brain phenomena have been misinterpreted. Pinto et al. have argued that the awareness that the subject demonstrates of stimuli presented to any part of the visual field is the same regardless of how they report that awareness (i.e. via the left hand, right hand or verbally), even though how information is processed varies depending on which hemisphere is presented with the stimuli (Pinto et al. 2017: 1235). If this is true, then Sperry's split-consciousness interpretation of split-brain phenomena would not be supported by the evidence.

Nonetheless, what we are interested in here is the process of science including the interpretations that have been held by scientists and so the split-consciousness interpretation remains useful to examine. Indeed, Pinto et al.'s dispute has been over whether participants

are consciously aware of the information processing, which they established by asking participants how confident they are with their answers (ibid.: 1233). Prior to the completion of the research, it was possible that participants would have responded in a different fashion, in which case the researchers could not have used their results to oppose Sperry's interpretation that each hemisphere was responsible for a consciousness that was not aware of certain information processed by the other. As such, the split-consciousness interpretation still remains something that many scientists would accept provided certain results were found that met the right criteria (e.g. that the self-reported confidence in information being processed varied depending on which hemisphere was responsible for responding). The earlier finding as reported by Sperry and others is thus still important for us to consider in assessing how we should interpret the science of consciousness and what constitutes reportable conscious behaviour.

Even if split-brain cases were not problematic, there are other cases such as that of the locked-in syndrome sufferer, who is incapable of almost any form of voluntary action. Sometimes such a person is capable of blinking or moving one eye (Damasio 2010: 234) but in total locked-in cases an individual cannot even do that (Bauer, Gerstenbrand & Rumpl 1979: 84). In the former case, we are likely to regard blinking as a form of reportability if it is done so in response to stimuli such as asked questions. In the latter case, we could only attribute consciousness to the individual via bootstrapping; we would have to know that they were in a physiological state that, under other circumstances, was accompanied by consciousness.

There is a limitation with attempting to use bootstrapping methodology in this way. If a particular physiological state is associated with a particular conscious state in an ordinary subject, then it will be by virtue of that state's ability to produce some form of recognisable conscious behaviour, and the individual suffering from total locked-in syndrome cannot possibly be in exactly this physiological state given that they lack the capacity to produce this behaviour.

Here I must reiterate the fact that, if we can observe correlations between conscious states and physiological states, any such observation of a physiological state that accompanies a conscious state must be an observation of a physiological state that is responsible for the production of recognisable conscious behaviour. Contrary to this claim, we might wish to hypothesise, for example, that regarding a physiological state P, which is present when we report being in conscious state C, only a part of P is responsible for the conscious state. This is what we would have to claim if we were to say that somebody incapable of producing recognisable conscious behaviour nonetheless has sufficient physiological conditions for a conscious state. We might wish to posit, for instance, that there are two parts to physiological state P – there is Pa, which is responsible for my being in C, and Pb, which is responsible for the capacity to behaviourally indicate that I am in C. If this were so, we would be able to remove Pb without removing Pa, and thus remove the capacity to behaviourally indicate the presence of a conscious state without removing the conscious state.

Nonetheless, it is not possible to experimentally differentiate between this possibility and the possibility that, by removing Pb, we have also removed the physiological state sufficient for my conscious state. If Pa is also present when we do not report being in conscious state C, we would have no reason to regard it as being a correlate of C at all; our observations would have given us no reason to implicate that part of P as opposed to the overall physiological state P.

Total locked-in syndrome may appear to be a clear counterexample to this, but this is not so. If a total locked-in syndrome sufferer did have a physiological state resembling that of somebody exhibiting recognisable conscious behaviour with simply the absence of the physiological states required to generate such behaviour, we would be just as justified in claiming that the physiological states required to generate such behaviour *are* essential for the presence of conscious states, and thus that the total locked-in syndrome sufferer lacks conscious states.

The alternative presents itself only in the situation where the total locked-in syndrome sufferers were to recover sufficiently from their locked-in state to report having been conscious the entire time. Nonetheless, we would then have to concede that, while they were in the locked-in state, they exhibited no recognisable conscious behaviour but were nonetheless conscious. This would make the notion of recognisable conscious behaviour a somewhat vague criterion; we would have to assume that we know that a state is conscious if and only if it produces such behaviour, and that a state is reportable if and only if, given the correct set of conditions (such as that they successfully recover from locked-in syndrome), they are able to produce such behaviour. The trouble with this is that there are likely a great number of physical states that could produce such behaviour given the correct set of conditions. This notion of recognisable conscious behaviour would be about all states that are hypothetically able to produce reports *in some situation*.

Scientists have broadly seemed aware of issues related to systems that may be conscious but whose consciousness we have no knowledge of, but have often wished to avoid discussing them in the preliminary stages of scientific investigation:

> It is not profitable at this stage to argue about whether “lower” animals, such as octopus, fruit flies, or nematodes, are conscious. It is probable, however, that consciousness correlates to some extent with the degree of complexity of any nervous system. When we clearly understand, both in detail and in principle, what consciousness involves in humans, then will be the time to consider the problem of consciousness in lower animals.
>
> For the same reason I won't ask whether some parts of our own nervous system have a special, isolated, consciousness of their own. If you say, “Of course my spinal cord is conscious but it's not telling me,” I am not, at this stage, going to spend time arguing with you about it (Crick 1994: 21)

If we assume that disregarding the conscious states of any physiological state or system that cannot produce reports is conducive to clear scientific research, then we can simply focus on those cases where reports are produced, and, in order to avoid any ambiguity, the physiological states common to sufferers of total locked-in syndrome will be considered cases of the presence of consciousness only if there have been cases where recoverers have produced reports about their awareness. As such, we will still be using the criterion of a state producing recognisable conscious behaviour as the basis for our observations of conscious states.

It is because of ambiguous cases such as the split-brain and blindsight that Baars wished to avoid discussing anything other than clear cases where a subject can voluntarily report being in a particular conscious state (Baars 1988: 12). This is the reason we have aimed to describe only PC states here; these form the basis for the scientific study of consciousness.

[2.5] Distinguishing the presence of a PC state from its absence

The complication that most clearly threatens the current picture of a science of consciousness arises when we consider how it is that scientists can distinguish the presence of PC states from their absence. As already discussed, this distinction is essential if we are to determine what the observable correlates of PC states are. The criterion Baars gave for the identification of the absence of such states was that they must not be reportable even under optimal conditions (Baars 1988: 12). What optimal conditions are seems difficult to determine and no explanation has been forthcoming. The physiological states of an individual with total locked-in syndrome would not be considered to be PC states if it was not for the fact that under certain conditions their physiological states were capable of producing recognisable conscious behaviour such as reports.

Many states that are not PC states, such as those of an individual under general anaesthesia, somebody having an absence seizure, or somebody in a certain stage of sleep, are considered such because the individual never exhibits any recognisable conscious behaviour while in these states nor are they able to produce reports about them after entering a recognised conventional state of consciousness. Whether these states would be reportable under optimal conditions is highly dependent upon what those optimal conditions are, and it is not clear that there is *no* condition under which an individual under general anaesthesia could produce recognisable conscious behaviour.

As explored in the previous section, there are some states for which it is more difficult than others to determine whether what they are exhibiting could be considered recognisable conscious behaviour. This did not prove to be a problem for identifying PC states; while we may not have known at one stage that total locked-in syndrome was an example of a state capable of producing recognisable conscious behaviour under the right conditions, once we did make this discovery we could classify total locked-in syndrome as a case of consciousness in the absence of those recognisable conscious behaviours (unless the patient recovers sufficiently). Nonetheless, there are problems caused when we attempt to use the inverse of this criterion to identify the absence of a PC state.

Verbal reports of conscious states, for instance, do constitute recognisable conscious behaviour, but it is common for scientists to consider these to not be requirements for consciousness. Crick stated that "a language system (of the type found in humans) is not essential for consciousness – that is, one can have the key features of consciousness without language" (Crick 1994: 21) and Ramachandran stated that he had "difficulty accepting that [qualia] requires a fully-fledged language in the sense that we usually understand that term" (Ramachadran 2004: 150). Whether we consider global availability for arbitrary response to be a criterion for identifying the presence of PC states seems to depend on whether such global availability can account for our ability to produce recognisable conscious behaviours, and it seems as though it can but with important caveats.

The most important caveat is that global availability does not necessarily produce recognisable conscious behaviours; a total locked-in patient may exhibit no such behaviours and never recover from that state. While global availability may be necessary for such behaviours, it is not sufficient. If somebody is to recover from total locked-in syndrome, and the other faculties required for producing those behaviours begin functioning in the way they did prior to them entering the locked-in state, they would then be able to report those states that were globally available at the time of them being in a locked-in state. We could tell a similar story about cognitive faculties that are necessary, but do not suffice, for global availability. For instance, the ability to distinguish between different stimuli is required for many globally available responses but is not sufficient for global availability, since such may simply be performed by an isolated system. As such, it seems arbitrary to draw the line at global availability as sufficient for conscious states; just as global availability requires certain physiological or cognitive conditions to produce recognisable conscious behaviour, so does the ability to distinguish between different stimuli. Indeed, it seems difficult to conceive of some physiological or cognitive (or indeed physical) state that we could judge as not being able to produce recognisable conscious behaviour under *some* circumstances.

As an example, let us imagine that if physiological state Q is present, I will exhibit the recognisable conscious behaviour of being in pain. It may be the case that being in a part of that overall physiological state, Qa, suffices for me being in pain, whereas another part, Qb, is required for me to exhibit pain behaviour, and so if I am under general anaesthetic but by an unfortunate accident remain entirely conscious I will remain in Qa but not Qb, feeling pain but unable to express it. We might then say that Qa *under optimal conditions* will produce recognisable pain behaviour. However, if I take another system consisting of a very advanced robot that exhibits the same recognisable conscious behaviour only when I shine a beam of light into its sensor, would we similarly wish to state that the beam of light *under optimal conditions* will produce recognisable pain behaviour and thus that something being a beam of light suffices for the presence of the conscious state of being in pain? To argue that there is a

difference between the two cases in terms of the presence of pain we would have to make a distinction between the sort of role Qa plays in determining how pain behaviour happens and the sort of role the beam of light plays, but this distinction must be made on the basis of knowing which sort of role suffices for the production of consciousness.

We might wish to argue that Qa plays a comparatively more complex role to a beam of light. Indeed, we could say, the beam of light simply activates another mechanism that then does all the work of generating a response, which is not the case for Qa. This argument would be inadequate. Contrasting the role of Qa and the beam of light assumes that we have already established what level of complexity or what sort of role a conscious state must have in the production of recognisable conscious behaviour, beyond it simply being a necessary component of the production of such behaviour, which is precisely what a science of consciousness is supposed to be able to establish.

The trouble here is that the vagueness of the notion of "optimal conditions" has infected the notion of something's being able to produce recognisable conscious behaviour. A mere beam of light can produce recognisable conscious behaviour under the correct conditions, although obviously not alone. The question is how involved in the production of that behaviour a physical, or physiological, or cognitive state has to be in order to suffice for the presence of a conscious state.

As another example, we could attempt to state, as Crick did, that memory is essential for consciousness, on the grounds that a state will never be able to produce recognisable conscious behaviours without memory. Somebody with epileptic automatism for instance is not able to report their conscious states upon recovering from an absence seizure and so there is good justification for assuming that their physiological state at the time was not such that it enabled them to retain information about being in that state. Nonetheless, if we are again going to begin simply with the criteria that something is a PC state if it can produce recognisable conscious behaviour under certain optimal conditions, or in the correct physical context, then it is not clear that even an individual having an absence seizure is not in a

conscious state. Presumably the physical effects of the physiological states in epileptic automatism, even if not stored in memory, still exist in the form of changes, however subtle, to the physiological state of the individual and the world surrounding her. As such, it is presumably physically possible for those physical effects to be used as input into a sufficiently advanced mechanism that can produce recognisable conscious behaviours such as verbal reportability as a result of that input. A mechanism is presumably physically possible that, if integrated into an individual's physiology in a certain way, could allow them to awaken from an absence seizure and say, "Oh, that seizure felt strange! I was walking around and looking at things with a weird tingly sensation running through my head." Indeed, there is nothing in principle impossible about any physical state being capable of producing any recognisable conscious behaviour if information about that state were accessible to a system capable of producing a particular kind of recognisable conscious behaviour based on their input. This being so, there is no physical state that could not produce every possible kind of recognisable conscious behaviour under certain "optimal" conditions if this is not defined, and thus every physical state could produce every recognisable conscious behaviour under optimal conditions if "optimal conditions" are not defined.

Defining these conditions is not a simple task. In fact, it is an impossible one. We might wish to state, for instance, that global availability suffices for consciousness because it determines the overall character of our responses to a much greater degree than, say, a neurological system designed to detect only a small number of stimuli. The trouble is justifying why determining the character of our responses to that degree is more likely to be sufficient for consciousness than detecting only a small number of stimuli. If we were to attempt to justify the distinction between these conditions, this would need to either be performed using *a priori* justifications or through empirical investigation, neither of which seems adequate. Even if we could determine *a priori* that conscious states were always accompanied by neurophysiological states, or cognitive states, it is presumably not possible for us to determine *a priori* which specific examples of these correlate with which specific conscious states any more than it is

possible to determine *a priori* which states are responsible for our ability to convert food into energy. Recognisable conscious behaviour is produced by certain mechanisms producing certain effects, and to suppose that we can know *a priori* what these mechanisms are is not only implausible but would also be to suppose that a science of consciousness could be conducted wholly *a priori* and thus would not meet the basic criterion we set out for something being a science in the first place, which was that it must be able to differentiate between possibilities on the basis of observation.

As such, we must rely on empirical investigation to determine what constitute "optimal conditions." Yet this is extremely problematic because all we could tell from empirical investigation is which physiological or cognitive states produce recognisable conscious behaviours under which conditions. We have no measure of which of these states are "optimal." Even if we have a full list of all physiological and cognitive states that produce recognisable conscious behaviour under all conditions, we will have to consider some items on the list "non-optimal" and others "optimal," but it is difficult to see how this could be based on the presence or absence of some essential aspect of conscious states and not on arbitrary criteria. For instance, if I am to correctly claim that I am in pain, this must be true by virtue of the presence of certain states. If there are a variety of situations within which I am unable to tell if those states are present via observation, categorising some of those situations as "optimal" and "non-optimal" does not resolve this issue; I am still unable to tell whether those states are present in some non-optimal conditions as well as optimal ones. If we were to determine that a state was optimal only if it required the use of physiological or cognitive systems that the person had when they were healthy, for instance, we would have to present a case for why we have used this criteria. It could not be for *a priori* reasons that we consider only such states to be conscious, for the same reasons explained above, and empirical observation would only reveal the various states that we have variously determined to be in optimal and non-optimal conditions rather than giving us justification for the distinction between those two categories.

The notion of what constitutes a state being capable of producing recognisable conscious behaviour is thus a vague one, and it is difficult to see how it could be otherwise. This is extremely problematic because we require a notion of something being *incapable* of producing recognisable conscious behaviour if we are to state that such cases are examples of the absence of conscious states. Without such cases, we have no examples of the absence of conscious states and a contrastive approach, such as that required for current consciousness science to make progress on its task, is impossible.

[2.6] Summary

Over the course of this chapter, we have established that observation should play a role in determining the truth of claims about the subject matter of science however "observation" is understood within the context of that scientific practice [2.1]. It was established that consciousness science operates by contrasting the presence of consciousness with its absence [2.2] – [2.3]. We explored how this contrast is possible by reference to the capacity to produce recognisable conscious behaviour [2.4], and we found a deep problem in defining what should constitute this capacity, such that it seems impossible to distinguish between something with this capacity and something without it [2.5].

In chapter 3, we will press the issue in order to establish that it is not simply the ambiguity in what constitutes the capacity to produce recognisable conscious behaviour that makes it such that we cannot distinguish the presence from the absence of PC states, but rather it is something inherent in the very idea of using empirical observations to make such a contrast.

3

Unrecognisable Consciousnesses

[3.1] Conscious states may fail to produce recognisable conscious behaviour

Even if there was no problem related to the ambiguity of what it means for something to produce recognisable conscious behaviour such that we could easily distinguish between those states that are capable of doing so and those that are not, there is a further serious issue with assuming that even PC states should necessarily be capable of producing recognisable conscious behaviour in the first place. This stems from the fact that a PC state is simply a conscious state that we can indicate the presence of through reports or behaviour; it does not follow that this (numerically) same conscious state could not have been present without that capacity to provide a recognisable communication about its presence. If I have a certain conscious state by virtue of being in a certain neurophysiological state, then even if that conscious state is only possible to communicate by virtue of some further neurophysiological conditions it does not follow that the neurophysiological state underpinning the conscious state is necessary linked to that underpinning the capacity to communicate such that they are not in principle separable. Certainly, it is not inconceivable that *any* physical state could be accompanied by consciousness whether or not it is even remotely involved in the production of recognisable conscious behaviour. Indeed, the fact that it is conceivable that there is consciousness where there is no recognisable conscious behaviour is a prerequisite of panpsychism, which, according to Skrbina, has been one of the (if not the) most widely held conception of the nature of mind across history (Skrbina 2009: 2). The question that is raised here is whether we should assume that conscious states are necessarily linked to the capacity to produce recognisable conscious behaviour.

This seems to go against our entire reasoning as to why we were studying PC states to begin with. However, the ambiguity cannot be avoided; for it to be true that you are in a paradigmatic conscious state of feeling a certain sensation, or perceiving a certain scene, it must be true that you are in a conscious state of feeling that sensation or perceiving that scene and that you can indicate that this is so. We have been supposing that the two are linked up until this point but whether this is so is a legitimate question.

There are, nonetheless, reasonable objections we could have to the suggestion that conscious states could exist without recognisable conscious behaviour. I will first answer this question with regard to consciousness in general. I will ascertain whether it is plausible that consciousness could exist in the absence of recognisable conscious behaviour, and indeed whether there is any justification for claiming that any particular state is an example of the absence of consciousness.

Secondly, I will answer this question with regard to specific conscious states. I will ascertain whether it is plausible that a particular conscious state could exist even in a subject's body where that subject reports having no awareness of it, and whether there is any justification for claiming that any particular state is an example of the absence of that particular conscious state.

Before beginning either of these tasks, I will clarify one thing; when I refer to a conscious state that produces no recognisable conscious behaviour I am not referring to a necessarily empirically unobservable state, but rather an empirically *unidentifiable* state. It could be the case, for instance, that the physiological state of somebody under general anaesthetic is accompanied by a conscious state that does not produce recognisable conscious behaviour. This does not require the assumption that conscious states are identical to physiological states although it would be very simple to argue that such states would be observable if we did make such an assumption. It is certainly possible, then, for us to argue that conscious states are observable by virtue of us being able to observe them qua physiological states if we take it to be true that conscious states can be reduced to physical

states. Even if we do not believe conscious states to be empirically observable, we have already established that they must be able to produce empirically observable effects. In the case of somebody under general anaesthetic exhibiting no recognisable conscious behaviour, there are still observable effects of their physiological state, but we do not recognise these effects as a consequence of the presence of any conscious state. The problem is that, if somebody under general anaesthetic is feeling intense pain, we cannot identify any effects as resulting from that pain, but this does not entail that the pain has no observable effects any more than the fact that somebody severely lacking empathy may be unable to identify your sadness from your tears entails that your sadness has no observable effects. In such a case, we would assume that your sadness is responsible for your tears but that, even though the unempathetic person may be unable to identify that sadness, they are still observing its effects. Similarly, our inability to recognise that certain observable states, such as the physiological states of somebody under general anaesthetic, are evidence of the presence of conscious states may not entail that they are not so.

The distinction is crucial because the idea that a consciousness could exist that produces no observable effects is trivial and non-interesting in a rather obvious sense. The same could be said for any phenomena; there could be chairs and buildings that exist but produce no observable effects, but this conceivable possibility tells us nothing about chairs and buildings. For us to simply be incapable of identifying whether consciousness is present even in some cases where it produces observable effects, on the other hand, implies an epistemic limitation; the tools we are using to understand consciousness – in this case a proposed scientific study of consciousness – are incapable of doing the job we have given them. It is this limitation that I wish to demonstrate.

[3.2] Assessing possible cases of consciousness that cannot produce recognisable conscious behaviour

To determine that it is entirely plausible that there are possible cases of consciousness that cannot produce recognisable conscious behaviour, I will first need to present such speculative cases and then ascertain three things.

The first is whether it is even a genuine possibility that such cases exist.

The second is, even if it is possible, how likely such a case would actually be. It is possible for a car to be in any unobserved car-shaped and car-sized space, but without there being any further reason for thinking there would be, I will almost always be wrong if I point to such a space randomly and claim that there is one there. The same rule should apply to consciousness; if I have no reason for thinking there will be consciousness in a particular observed state, there is a case to be made that it is plausible for me to assume that there is no consciousness present.

The third is, even if it is possible and not unlikely for a state of consciousness to be in an observed place, whether we are holding a science of consciousness to an unreasonable standard if we assume that it should be able to account for such a possibility. It is possible that lions have wings and lay eggs but that they also interfere with our senses in such a way that they appear to us to give birth to live young and move around on foot, but we would not expect this possibility to affect the claims that scientists may wish to make, such as that lions are mammals and cannot fly. As such, it may seem unreasonable to hold a science of consciousness to this standard and expect it to be able to account for possibilities that no observation could give us information about.

[3.2.1] Speculative cases of consciousnesses that cannot produce recognisable conscious behaviour

Given that these are speculative cases, there is no limit to an observable state that can be posited as being accompanied by a consciousness. The pen on my desk could be accompanied by consciousness. The wind blowing a leaf down the street could be accompanied by consciousness. The fact that the observable effects of such states have very little similarity to the effects of our consciousness would tell us nothing if there were consciousness present in such cases. Indeed, even if consciousness is not present in such cases, the lack of similarity between those effects and recognisable conscious behaviour does not tell us this fact unless we first assume that consciousness *must* produce recognisable conscious behaviour.

To use the example of being administered general anaesthetic, we are incapable of producing any recognisable conscious behaviour during or indeed following being in a state induced by general anaesthetic, unless we are unfortunate enough to be an individual who is awake but unable to speak or move until after the procedure. However, even if we are not such an individual and upon waking we report having no recollection of experiencing anything during our procedure, it does not necessarily follow that we had no experience at the time; it is certainly conceivable that the faculties disabled by general anaesthetic are not those responsible for the presence of pain but are those responsible for the presence of our capacity to produce recognisable pain behaviour. As such, we may be able to feel pain under general anaesthetic even if we lack the capacity to flinch away from it, report feeling it, or recall feeling it after the anaesthetic wears off.

A denial of this requires us to assume that the capacity to produce such behavioural attributes either must, or at least is likely, to accompany all instances of consciousness. There are several cases to be made for such a denial, but all of them are demonstrably flawed.

[3.2.2] The denial that such consciousnesses even could exist

There has long been a tradition in philosophy to deny any legitimacy to claims that there exist entities or features of the world that cannot be verified through observation. I will refer to this tradition broadly as "verificationism". This tradition has roots in the philosophy of Hume, who claimed that if we come across any work of metaphysics unrelated either to numerical, quantitative reasoning or experimental reasoning resulting from facts discovered through experience, we should "Commit it then to the flames: For it can contain nothing but sophistry and illusion" (Hume 1748: E12.34). The opposition was to "profound and abstract philosophy" that is not only "painful and fatiguing, but… the inevitable source of uncertainty and error" (ibid.: E1.11). Over the last century, there are other examples. Wittgenstein attacked the notion that we can possibly refer to something that is not intersubjectively observable, such as in his "beetle in the box" analogy, where he stated that the belief that we know what our own pain is from private experience is like everybody using the word "beetle" to describe what they have in a private box and then assuming that the word "beetle" now means whatever you see in your box. The trouble is that others may be using the same word to describe something else, or nothing at all, and so "The thing in the box has no place in the language game at all; not even as a *something*" (Wittgenstein 1953: 293).

Ayer rejected metaphysics wholesale:

> With regard to the relationship of philosophy and the empirical sciences, we have remarked that philosophy does not in any way compete with the sciences. It does not make any speculative assertions which could conflict with the speculative assertions of science, nor does it profess to venture into fields which lie beyond the scope of scientific investigation. Only the metaphysician, does that, and produces nonsense as a result. (Ayer 1936: 167)

Another philosopher I will place in the same category is Carnap, despite his preference of the term “confirmation” to the term “verification” (Carnap 1936: 420) to describe how we determine the truth value of claims, since he believed that “no (synthetic) sentence is ever verifiable” (ibid.: 420) and is rather only more and more confirmable over successive observations (ibid.: 423). For a statement to be confirmable, for Carnap, was for there to be some possible observation that would confirm that sentence (ibid.: 457). Although he cautioned against claims to the effect that unverifiable claims are meaningless, since this would also deny meaning to many scientific sentences (ibid.: 421), I still categorise him along with other verificationists for the purposes of this section purely because of his refusal to allow legitimacy to any questions that experience will not allow us to answer:

> A (pseudo) statement which cannot in principle be supported by an experience, and which therefore does not have any factual content would not express any conceivable state of affairs and therefore would not be a statement, but only a conglomeration of meaningless marks or noises. (Carnap 1928: 328)

The tradition is strong in the works of much more recent philosophers, with Ladyman, Ross and their collaborators claiming that “no hypothesis that the approximately consensual current scientific picture declares to be beyond our capacity to investigate should be taken seriously” (Ladyman, Ross et al. 2007: 29).

To philosophers of this general tradition, it will likely be seen as an egregious example of speculative metaphysics to posit that there may be examples of conscious states that no observation will allow us to identify. Rather, it could at least be argued that I would need to have at my disposal some argument as to why verificationism, broadly construed, is false before I can claim to legitimately be able to refer to such conscious states as real possibilities.

The criticism could run as follows. I am arguing that there is some truth value to a claim such as this: “There are conscious states that do not produce recognisable conscious behaviour.” For a verificationist, the claim that there exist states that do not produce evidence of themselves should at best not be taken seriously and at worst be regarded as meaningless, since it is clear that there are no conditions under which we could verify the existence of such states. If such claims are thus to be ignored, then there is no argument to the effect that they can cause a problem for the science of consciousness.

The trouble with this criticism is that it is not enough for us to ignore the above statement; rather, we must know that its *negation* is true. We must know that the statement, “All conscious states are capable of producing recognisable conscious behaviour,” is true. Yet, if my claim above could not be confirmed, it is difficult to see how its negation could be. The conditions under which we could confirm this second claim would have to be conditions under which we observe conscious states occurring alongside recognisable conscious behaviour, but it is not conceivable that we could observe a conscious state that does not occur alongside recognisable conscious behaviour, nor is it conceivable that we could observe recognisable conscious behaviour and confirm that this behaviour is present in the absence of a conscious state. Thus we can know ahead of any investigation that all observations will support this conclusion, since no observation to the contrary is conceivable. On the other side of the coin, we could not confirm that the above statement is false by relying on any possible observation of recognisable conscious behaviour. These considerations show that the above statement could not be supported by induction and could only be supported by analysis of the rest of the sentence, and so must be deducible from the meaning of the expression “capable of producing recognisable conscious behaviour.”

The only way it could be deduced from the definition of the terms “capable of producing recognisable conscious behaviour” that *all* conscious states are so capable is if the phrase “conscious state” or “consciousness” simply means “whichever state is capable of producing

recognisable conscious behaviour." In such a case, the claim would be tautologically true, rather like claiming that a triangle must have three sides.

If this is so, then the only thing that needs to be explained by a science of consciousness is how we have the capacity we do to produce recognisable conscious behaviour, which falls within the research project of analytical behaviourism. This behaviourism, which is often attributed to Ryle, attempted to describe mental states purely in terms of a person's behaviour:

> Overt intelligent performances are not clues to the workings of minds; they are those workings. Boswell described Johnson's mind when he described how he wrote, talked, ate, fidgeted and fumed. (Ryle 1949: 58)

To explain how we can have mental states that do not produce such behaviours, such as if we have a private thought or disguise our feelings, Ryle identified the mental state not with the behaviour itself but with the disposition to produce that behaviour. For Ryle, to be in pain is to be disposed to shout out and hop around, or to avoid touching that same object again, or to jerk your hand away, or to clutch a certain area of your body. To understand something is to be disposed to respond in certain ways when asked certain questions, or to have "correctly drawn further consequences from different stages of the argument and indicated points where the theory was inconsistent with other theories" (ibid.: 170). He summarised:

> In short it is part of the meaning of 'you understood it' that you could have done so and so and would have done it, if such and such, and the test of whether you understood it is a range of performances satisfying the apodoses of these general hypothetical statements. (ibid.: 170)

The main problem behaviourism had to contend with was that reducing beliefs (for example) to behavioural dispositions "offers us no prospect of analysing talk of beliefs in entirely non-mentalistic terms" (Smith & Jones 1986: 145). A disposition to act in a certain way must be defined with reference to other mental states and thus other dispositions. Being in pain might be a disposition to hop around yelling, but if I'm in a situation where doing so would cause me significant social embarrassment, I might just stifle my suffering instead and have to be content with perhaps making a face and a quiet remark. As such, that state of pain I was in would have to be defined as a disposition to act in a certain way *provided* I am not in another mental state that involves inhibiting that behaviour, such as that of fearing embarrassment. This means that I must define a mental state with essential reference to another mental state, which I also must define according to a certain disposition. A disposition to act a certain way thus must be defined by reference to other dispositions, which must then be defined by reference to other dispositions and so on.

The trouble is that a disposition can presumably only be realised by a certain state that produces that behaviour. However, if that state does not *necessarily* produce such behaviour, and it can be suppressed or modified, then this seems to suggest that the state that would usually produce my pain behaviour – my state of being in pain – can be present in the absence of any particular behaviour. At best, if there is a state that sometimes produces pain behaviour and is also present where I am stifling my pain then it seems unusual not to define my pain according to that state but rather according to the behaviour that it may or may not produce.

A similar effort to identify mental states with the disposition to produce certain behaviours has been revived as the research project of "illusionism", which makes sense given that Dennett, who has been considered the most prominent defender of illusionism (Frankish 2016: 12), was once a student of Ryle's. Whether illusionism as a whole could be considered to be a verificationist project or not, perhaps its most prominent advocate has accepted a link between verificationism and his own work. In response to a claim by Rorty that Dennett could not avoid "becoming the Village Verificationist" (Rorty 1982: 342–343), Dennett once stated

that he was “ready to come out of the closet as some sort of verificationist, but not, please, a Village Verificationist,” preferring instead to be considered an “*Urbane* Verificationist” (Dennett 1982: 355).

The illusionist seeks to explain our “reports, judgements and intuitions about our own consciousness” (ibid.: 37) rather than the properties or qualities that we purport to describe when we talk about our conscious experience. The focus here, if we are truly embarking on a verificationist-friendly project, is to identify that which is responsible for the empirically observable aspects of these judgements and intuitions, such as our behaviour and descriptions.

If, though, it is consciousness that we are claiming to study, it will not do to simply identify our claims about consciousness; we would also have to suppose that those claims could have a truth value. If my claim, “I am conscious now and am experiencing happiness,” could not be true or false then we could not claim to be studying consciousness at all. To say that claims about consciousness have no truth-value entails that consciousness cannot be studied in the same way that to say that claims about the afterlife have no truth value entails that being a medium is not a legitimate profession.

If we admit that they do have truth value, it is difficult to see what this truth value consists in once we accept that a conscious state is simply whichever state is capable of producing, for instance, claims that one is in that conscious state. There is presumably a certain set of physiological and cognitive conditions by virtue of which we make true claims about being in a certain conscious state, but there is presumably also a certain set of physiological and cognitive conditions by virtue of which we make false claims about being in a certain conscious state. To distinguish between the two, we would have to know which claims are accompanied by the conscious state being referred to and which claims are not, which is precisely the question that our science is supposed to be able to answer.

For example, two people may claim to be perceiving redness. One of them may genuinely be in the conscious state of perceiving redness, while the other may be lying to

disguise their colour blindness. In this case, if we believe that conscious states really exist (as we must if we believe in a science of consciousness) we must also believe that it is by virtue of the presence and absence respectively of a conscious state of perceiving redness that the first person is making a true claim and that the second person is making a false claim. The trouble is that, without knowing via some other means that one is true and the other is false, there is no method by which a science of consciousness can determine which is the case; it will simply be presented with one set of cognitive and physiological conditions underpinning one report of the presence of the conscious state of perceiving redness and a different set underpinning the other.

The issue is that we have two sets of conditions producing the same claim about the presence of a conscious state, but only one of those sets of conditions is truly accompanied by the state that is claimed to be present. If we simply define a conscious state as "the state responsible for the true claim," then a conscious state is "the state responsible for making a true claim about the presence of itself," which is not verifiable. All we can observe are two sets of conditions producing a claim, and, unless we have some independent means of observing the presence of the conscious state (which we cannot if we have so defined it) there is no way to distinguish between true and false claims regarding the presence of conscious states and a science of consciousness is thus impossible.

Moreover, the entire exercise of attempting to determine the difference between the physiological and cognitive states in these two conditions results from the assumption that there even is an empirically useful distinction between true and false claims about the presence of conscious states. This is problematic for two reasons:

1) The fact that we make certain claims does not entail that the things we are referring to are categories that will be useful once there is a scientific understanding of the mechanisms producing our claims. It could be the case that the claim, "I am in pain," corresponds to a variety of empirically observable states that there is no scientific use

in categorising together and that are sometimes present when no such claim is made. This is essentially the claim made by Paul Churchland regarding propositional attitudes, which was that there is little prospect of beliefs, hopes, and desires being reduced to, for instance, neuroscientific concepts (Churchland 1981).

2) If we are simply trying to correlate certain behaviour, such as pain behaviour, and the physiological causes of that behaviour, there is no interesting or metaphysically significant difference between the outright denial of the existence of consciousness and the reduction of consciousness to these physiological causes.

To expand on (2), if we are to reduce consciousness to physiological causes, then the term "consciousness" can only have a shorthand usage where it essentially means, "whichever physiological states produce such-and-such abilities," which, given that this is not how many other philosophers use the term, it would be just as congenial to their research goals to deny that there is such a thing as consciousness at all and to assert that *instead* of consciousness producing these effects, it must be physiological states. This move would not be available to somebody wishing to claim that there are phenomenal properties that cannot be defined purely in terms of their effects, because this would entail that when an individual claims to be in a conscious state, those properties either genuinely are there or they are not there, but that this distinction would not be observable via third-person scientific methodology since phenomenal concepts must be grasped first-hand (Stoljar 2005: 471). As such, to provide a verificationist defence of the science of consciousness we would have to deny that there are phenomenal properties, like the illusionist, but then we would also have to accept that the term "consciousness" is shorthand for a category of physiological states, which also means that we know ahead of time that the term can eventually be scrapped in favour of references to more specific physiological states. Given (1), though, they cannot claim to be interested in consciousness as a concept to apply to the empirical sciences because it may simply fail to line up with any empirically useful categories. There is no guarantee that consciousness terms

will line up with empirically observable states at all. As Paul Churchland stated in defence of eliminative materialism:

> The identity theorist optimistically expects that folk psychology will be smoothly *reduced* by completed neuroscience, and its ontology preserved by dint of transtheoretic identities… folk psychology is a radically inadequate account of our internal activities, too confused and too defective to win survival through intertheoretic reduction. (Churchland 1981: 72)

"Folk psychology" is a description of ourselves whereby we use "mentalistic" expressions such as "believes," "loves," or "is aware of," to explain our behaviour as opposed to expressions that find their origins in scientific discourse. Whether or not we accept that folk psychology has been shown to be so confused, if we are motivated by a verificationist philosophy, we would certainly be provided with a motivation to eliminate mental state terms if they were shown to have no bearing on our capacity to explain our behaviours and judgements in a way that corresponds to any verifiable states or processes, with the concepts of neuroscience being Churchland's chosen example.

Similarly, if the illusionist were consistent with her motivation for her position, which is that phenomenal qualities should not be posited since there is no clear way they can be implicated in the production of our claims and judgements about them, they should similarly deny the existence of consciousness since it too is not obviously implicated in the production of our claims and judgements about it. There is no clear category of phenomena to be explained here unless we believe in consciousness *prior* to empirical investigation, in which case the notion that we need to invoke the concept of consciousness to explain our capacity to make certain judgements may admit no empirically plausible explanation whatsoever.

The only way that the verificationist can defend her position is by assuming that "consciousness" refers wholly to an empirically observable category of physiological or cognitive states that are responsible for the production of our reports, but in this case the empirically observable category is being posited in order to account for consciousness, not the other way around. In this case, we would have an entirely consistent verificationist story of our physiological, cognitive and behavioural states if we were to abandon the idea of consciousness entirely; there would be nothing else left unexplained. As such, it is unclear how the verificationist position does not constitute a denial of the existence of consciousness rather than a defence of a scientific study of consciousness.

Since, given (1), consciousness as a concept is only useful insofar as it is a category of more straightforwardly empirically observable states, the foreseeable end result of such a verificationist project will be to scrap references to consciousness in favour of references to physiological and cognitive states. The only motivation that is consistent both with the illusionist project of explaining only the empirically observable processes responsible for our capacity to produce claims about conscious states and the refusal to abandon the term "consciousness," is the denial of non-scientistic metaphysics rather than the furthering of a scientific research project. The goal seems to be to place the bulk of the explanatory work for consciousness in the hands of scientists and scientistic philosophers rather than leaving it in the hands of philosophers who espouse a metaphysics that requires, or can receive, no input from science. While I have some sympathy for this goal – having stated that consciousness should lie within the scope of scientific investigation – the notion that we should just place the concept of consciousness in an explanatory framework for the production of our judgements about consciousness, and that much of this work is compatible with current research projects of neuroscientists and cognitive scientists, is plausibly detrimental to the development of a science of consciousness entirely if we were to assume that such a science possible. It is rather like claiming at Descartes' time that the immaterial soul could be understood through the life sciences in the form they existed in then. Clearly, the concepts related to minds needed

to be refined, as well as the practices and assumptions inherent in the life sciences, before we could see with sufficient clarity why such a study would have been doomed to failure. Similarly, the idea that consciousness, as currently understood, will neatly fit into the research projects of neuroscience seems easily liable to the same kind of mistake, and is at best a grand assumption about the future direction of both science and philosophy.

None of this constitutes a direct argument against verificationism, as we might instead just claim that consciousness does not exist. This claim will play a significant role in the later discussion about potential solutions to the Consciousness Science Paradox. But for now, it will suffice to state that a verificationist should not pin her hopes on a science of consciousness; it may serve as a useful tool in the dismissal of anti-scientistic metaphysics, but it fails on the grounds that it requires an ontological commitment to something that does not have a scientific basis.

To relate this back to the issue of conscious states that fail to produce recognisable conscious behaviour, the possibility that such states could exist cannot be denied on verificationist grounds, since such a denial would require that there already exists a scientific basis for the presence of consciousness. If we use such a denial then to prop up a science of consciousness, the justification will be entirely circular; we will know that all conscious states are verifiable because a science says so, and we will know that a science is possible because all conscious states are verifiable. The desire for an adequate justification thus can only be met if one of those two claims are justified another way. So, the denial of the possibility of conscious states that do not produce the only means we can identify them by – their recognisable conscious behaviour – does not leave us in a position to support the science of consciousness. If we wish to undermine the claim that there could exist conscious states that do not produce recognisable conscious behaviour, we must take another route.

[3.2.3] The likelihood that any given state is accompanied by a consciousness

If we wish to defend a science of consciousness, then, we might instead argue that, although there could be consciousnesses that do not produce recognisable conscious behaviour, it is generally safe to assume that, where there is no evidence of the presence of a consciousness, no consciousness is present. We might make this assumption in the same sense that we could quite safely assume that, if I am pointing at a randomly selected object and I do not know what it is (perhaps because I am blindfolded), I could guess that, whatever it is, it is not a lamp and presumably, given that the vast majority of objects are not lamps, I will usually be correct.

I assume some reasoning along these lines is what lies behind Crick's rejection of "special, isolated" consciousnesses, to repeat a section I quoted previously:

> I won't ask whether some parts of our own nervous system have a special, isolated, consciousness of their own. If you say, "Of course my spinal cord is conscious but it's not telling me," I am not, at this stage, going to spend time arguing with you about it (Crick 1994: 21)

The reasoning seems to be that such speculations are irrelevant, unhelpful, and unlikely to hit on the truth. Of course, what would be required for this reasoning to hold is for consciousnesses to be uncommon occurrences in nature, like lamps. The problem is in justifying this assumption.

To return to the example of general anaesthesia, if somebody is administered general anaesthetic and continues to be in a series of conscious states but exhibits none of the unified information processing that we might usually associate with the presence of consciousness, nor any of the particular neurophysiological states we associate with the presence of consciousness, the actual cognitive or neurophysiological processes that correspond to the

presence of consciousness would have to be comparatively *simpler* than those that are present in those we usually associate with consciousness. This is because, given that those cases where the presence of consciousness is identified are complex physiological and cognitive states, but that the states while under general anaesthetic are, if not less complex, certainly significantly different, the sufficient conditions for the presence of consciousness would have to be something in common with both kinds of physiological and cognitive state.

Now, that which is in common between two physiological or cognitive states cannot be *more* complex than either state; if there are two physiological states that are different, then only *part* of those states can be common between them. That part, being common to different physiological states, will thus more frequently be found in nature than either of the physiological states of which it is a part.

Even a state common to both an anaesthetised patient and a waking conscious person could be rarely found in nature. However, the likelihood of finding such a state would be higher if there are other physiological states unlike being under general anaesthesia and being in a standard state of wakefulness that are accompanied by consciousness. There could even be non-physiological states that could be conscious, in which case such states could be found with even greater frequency.

The likelihood of any given state being accompanied by consciousness can only be inferred from the knowledge of the sufficient conditions for the presence of consciousness. The problem is that we have no such knowledge once we allow for the possibility of consciousnesses that do not produce recognisable conscious behaviour. We have no way to eliminate anything even as simple as a subatomic particle as a candidate for the presence of consciousness if we give such an admission, since no observation gives us evidence by which we can eliminate the possibility that consciousness is present. Thus, we have no grounds to assume that consciousness is not present where there is no evidence of its presence. Absence of evidence is not evidence of absence.

Such an argument from improbability does not succeed then because it presupposes that conscious states necessarily accompany a particular kind of complex physical state. Whether or not this is so is something that can only be established following an account of the empirically observable states that accompany conscious states. As such, it cannot be used as a precondition for the possibility of such an account.

[3.2.4] Am I holding a science of consciousness to an unusually high standard?

The relationship between philosophical concerns and the progress of science is not like that of a trusted advisor to a policymaker; philosophers have for centuries offered criticisms of scientific practices and have even claimed that there is no rational basis for inductive reasoning (Hume 1739: 1.3.6) without this obviously leading to any significant change to the conduct of science, as observed by Whitehead:

> Science repudiates philosophy. In other words, it has never cared to justify its faith or to explain its meanings; and has remained blandly indifferent to its refutation by Hume. (Whitehead 1925: 16)

As such, the idea that a science of consciousness could be hostage to speculative possibilities of non-identifiable consciousnesses may seem to present an implausible and historically ignorant description of the relationship between science and metaphysics, where the scientist is unable to proceed until the metaphysician has determined whether there are problematic conceivable entities that would render her claims false.

It certainly seems that other scientific disciplines do not have to be attentive to concerns such as this. In claiming that a given chemical boils at a certain temperature at sea level, the chemist does not seem to be plagued with conceivable cases where the same chemical fails

to boil under the same conditions. The physicist does not have to worry that her claims about the strength of the gravitational force are incorrect because there are conceivable cases where objects composed of the same number and type of subatomic particles exhibit a much greater gravitational force. It perhaps seems strange then that I am attempting to discredit the science of consciousness on the grounds that there are conceivable consciousnesses that do not produce recognisable conscious behaviours.

The distinction between those other examples of general kinds of scientific work and the science of consciousness is not the standard to which I am holding them but rather their susceptibility to being discredited by claims regarding conceivable cases. A chemist's claims regarding the boiling temperature of a certain chemical will be unaffected by conceivable instances of something else occurring, since the chemist can simply restate that she is only interested in empirically observable cases or even that we can only say that something is a certain chemical if it does produce such effects. Even if she did not do so, the observable cases would still be a separate set of such instances to those conceivable counter-examples, and so we could easily restate all her claims about the boiling temperature of the specified chemical in terms that only refer to the observable cases.

These moves would fail with regard to the science of consciousness because facts regarding the empirically observable correlates of consciousness are dependent upon facts about unidentifiable correlates of consciousness; specifically, to state that we can derive from observation that a certain observable state is required for the presence of consciousness assumes that there are no unidentifiable states that are also accompanied by consciousness. If there were such states, and they were significantly different from those posited as being required for the presence of consciousness, this would render false all claims that those posited states are required for the presence of consciousness. This is not the case with regard to chemicals or objects exhibiting a particular gravitational pull; facts regarding any instances of these need not be dependent upon facts regarding other conceivable instances. It is not the case then that I am holding the science of consciousness to an unusually high standard.

Rather, it is the case that the science of consciousness is unusually susceptible to conceivable cases that would render its claims false. The claims that this science relies upon are non-verifiable because of their susceptibility to such conceivable cases and this entails that the truth of them cannot be determined by empirical observation.

[3.3] Assessing possible cases of conscious states that cannot produce recognisable conscious behaviour

All the concerns just mentioned in section [3.2] regarding consciousnesses failing to produce recognisable conscious behaviour can seemingly be sidestepped with a simple piece of reasoning, which goes as follows.

Even if a science of consciousness cannot tell us whether consciousness is present under many circumstances, it can at least tell us what the observable correlates are for the presence of particular conscious states. If a subject honestly reports being in pain under certain conditions but then under certain other conditions the same subject reports no pain, then it seems reasonable to assume that the conditions in the first set of conditions but not the second are responsible for the presence of pain. As such, even if a science of consciousness cannot tell us which conditions are sufficient for the presence of consciousness, it can tell us which conditions are sufficient for the presence of our particular conscious states. Given that this can presumably be done for all our conscious states, it is arguable that the entirety of the empirically observable correlates of our consciousness can be identified piece by piece through this scientific endeavour.

This reasoning has a straightforward appeal but requires that we make additional assumptions about consciousness than are apparent from the observable evidence alone, particularly regarding the number of consciousnesses present in a particular organism at any moment, and the unified nature of these consciousnesses.

What I am suggesting here is that when we claim that a conscious state is not present, it is possible that there is a separate consciousness present in that conscious state. When we are experiencing something, such as pain, and that experience ceases, it is possible that our consciousness has divided into two: the consciousness experiencing the pain and the consciousness no longer experiencing the pain. The purpose of these suggestions is not to make the case that there are indeed multiple consciousnesses in an individual human body, nor is it to make the claim that consciousness division is even possible, but it is to make the claim that a science of consciousness, while relying on the above statements being false, can neither demonstrate this through empirical means nor rely on the impossibility of such cases on the basis of reasons derived from *a priori* reasoning. This leads to the three main points that I will address. The first is how unknowable numbers of consciousnesses would render a scientific study of the correlates of conscious states impossible. The second is why we should accept, against many *a priori* objections, that it is even possible that there are unknowable numbers of consciousnesses present in an organism. The third is why, even if we have no *a priori* reason to believe in anything other than a singular, persisting consciousness per body, we should accept that it is a genuine possibility that this may not be the way things are rather than merely being a speculative case that is irrelevant to the existing science.

I will be assuming here that a consciousness need not produce recognisable conscious behaviour, for reasons that were outlined in [3.2]. The problematic nature of identifying the number of such consciousnesses will give another important reason that a science of conscious states is in fact impossible.

[3.3.1] The problem of an unknowable number of consciousnesses

If we were not to assume that a single organism (such as a human) had a single consciousness from birth to death, but rather had a multiplicity of consciousnesses, this would

create a significant problem for a science of consciousness. The problem would be fatal to such a science if we additionally avoided the assumption that a consciousness is unified over time.

With regard to the possibility of a multiplicity of consciousnesses, we would never be able to say with any level of certainty that a conscious state is absent simply because a subject reports it to be absent. If a subject says that he is not in pain, we will not be sure that another separate subject of the same body is not feeling that pain. As such, we will be unable to distinguish the presence of a conscious state from its absence in order to determine which empirically observable correlates are present in the former case but not the latter.

This problem seems at least partially avoidable if we argue that a science of conscious states could still, at the very least, tell us the empirically observable correlates of conscious states for a particular subject. One subject may claim not to be in pain and another in the same body may be experiencing pain, but if this is so then the correlates we have found are simply the empirically observable correlates of the former subject's consciousness. As such, there are still correlates to be found even if they are not the correlates required for *any* subject in that body to be in a particular conscious state.

This problem would cease to be avoidable if we do not grant the assumption that a consciousness must be unified over time. If it is possible for a consciousness to divide into multiple consciousnesses then finding empirical correlates of conscious states would be impossible. If a subject, for instance, claimed to be in pain at time *t*, while in physiological state *P*, and then were to cease to be in pain at time *t+1* while also ceasing to be in physiological state *P*, then we could only implicate physiological state *P* in that subject's state of pain if the subject did not divide into two consciousnesses, one of which was still in pain and one of which was not. In that case, the conscious state of being in pain could be said to have continued even while the (reporting) subject were to claim that it had ended. There would thus be no means for us to identify the empirically observable correlates of any conscious state without being able to rule out such a situation.

Such a situation may appear quite implausible. As it is, the possibility of the above situation *must* be accepted by a science of consciousness, for reasons that I will explain [3.2.4], after contending with the *a priori* arguments against situations involving multiplicities of consciousness, and consciousness divisions and combinations.

[3.3.2] *A priori* arguments against unknowable numbers of consciousnesses

There are three main kinds of *a priori* objection to the above cases of multiple consciousnesses that will be discussed here. The first are objections regarding the number of consciousnesses capable of inhabiting a single body, such as positions that suppose there could only be a single consciousness to a body. The second are objections regarding the possibility of consciousness dividing. The third are objections regarding the possibility of consciousness combining.

The first sort of objection can be attributed to P.F. Strawson, who argued that the concept of a mind is dependent upon the concept of a person. For Strawson, predicates such as "sensations, thoughts, feelings, perceptions," are things we attribute to human beings just as "physical position" is (Strawson 1974: 187) and that it is a Cartesian dualist approach to separate these references into those about consciousness and those about the bodily condition. The anti-Cartesian, instead, asserts that the notion of a mind or consciousness is dependent upon that of a person (ibid.: 188). The main problem with the Cartesian approach is that, for us to be able to coherently refer to consciousnesses we must be able to distinguish between one and multiple consciousnesses, but we have no criteria by which we can count consciousnesses; if a consciousness is an entity bearing a certain relation to a person then any number of consciousnesses could bear that relation to a single person (ibid.: 191). We could not necessarily distinguish them in terms of the character of their experiences since they might be qualitatively indistinguishable but it is not clear how else we could distinguish them

(ibid.: 192). As Quine stated, there should be clear identity criteria for any entity we wish to postulate; there can be “no entity without identity” (Quine 1969: 23).

If we were to adopt an anti-Cartesian approach, then, it would seem that we could not posit that a consciousness could divide into multiple consciousnesses within the same body because to do so would entail accepting that we can identify a consciousness beyond simply identifying a body that has the particular mental or conscious states we wish to account for.

The second sort of objection, that consciousnesses cannot divide, is perhaps most famously embodied in Kant’s argument for the transcendental unity of apperception. The argument states that consciousness must be unified because this is the only way to account for our representation of things existing across space and time. If the argument is accepted, we cannot allow for the possibility of consciousness dividing or combining over time, since this would entail disunity over time and thus would not be able to account for our representation of things persisting over time. As Kant stated in the B-deduction of the Critique of Pure Reason:

> The **I think** must **be able** to accompany all my representations; for otherwise something would be represented in me that could not be thought at all, which is as much as to say that the representation would either be impossible or else at least would be nothing for me. (Kant 1781: B132)

The entity that Kant refers to as the “I think” must be able to accompany all of my representations, including those persisting across time. Indeed, Kant wrote that the unity of consciousness is an essential precondition for all possible perception.

> The objective unity of all (empirical) consciousness in one consciousness (of original apperception) is . . . the necessary condition even of all possible perception. . . (ibid.: A123)

If we accept Kant's conclusion, there would be no way for us to make sense of the notion that we could have been a single consciousness at time A and then divided into two at the later time B, since to do so we would have to be capable of conceiving ourselves apart from the unity that existed at time A.

The third form of objection is related to the impossibility of conceiving how consciousnesses could, in principle combine. There is plenty of argument in the panpsychist literature that consciousnesses, experiences, or subjects, cannot combine to form one "larger" consciousness, experience, or subject. This problem is known as the "combination problem" and is attributed to William James, who wrote:

> Take a hundred [feelings], shuffle them and pack them as close together as you can (whatever that may mean); still each remains the same feeling it always was, shut in its own skin, windowless, ignorant of what the other feelings are and mean. There would be a hundred-and-first feeling there, if, when a group or series of such feelings were set up, a consciousness *belonging to the group as such* should emerge. And this 101st feeling would be a totally new fact; the 100 feelings might, by a curious physical law, be a signal for its *creation*, when they came together; but they would have no substantial identity with it, not it with them, and one could never deduce the one from the others, nor (in any intelligible sense) say that they *evolved* it. (James 1890: 160)

This is a description of the problem of combining separate feelings, whereas Goff described a deep intuition that "***subjects aren't combinable***" (Goff 2017: 171). He described the problem in more detail:

> Consider a physical ultimate that feels slightly pained, call it LITTLE PAIN 1. Consider ten such slightly pained ultimates, LITTLE PAIN 1, LITTLE PAIN 2, etc., coming together to constitute a severely pained macroscopic thing, call it BIG PAIN... But what it feels like to be LITTLE PAIN 1 is not part of what it feels like to be BIG PAIN. LITTLE PAIN 1 feels slightly pained, BIG PAIN does not. The phenomenal character of LITTLE PAIN 1's experience, i.e. feeling slightly pained, is no part of the phenomenal character of BIG PAIN'S experience, i.e. feeling severely pained (Goff 2006: 57-58).

The problem described is that there appear to be aspects of the character of conscious states that cannot be captured by reference to the addition of posited constituents of that overall state. This is particularly a problem for panpsychists because a consciousness such as human consciousness is taken to arise once many constituent parts, such as neurons, which themselves are composed of entities that are conscious, are placed together in a certain fashion. If panpsychism is to avoid conventional physicalist problems of consciousness then, arising from difficulties describing how conscious states can arise from non-conscious states, it must overcome the problem of explaining how consciousness can arise from the combination of other consciousnesses.

Nonetheless it is also seemingly a problem for the current criticism of consciousness science; if the posited combinations of consciousnesses cannot happen then we might think it odd to argue that composite consciousnesses could present a problem for any position. This accounts for three kinds of *a priori* objection to the idea of consciousnesses combining and dividing. We can object in the sense that Strawson objects, by denying that we can refer to individual consciousnesses within a single person, in the way that Kant objects, by asserting that consciousness must be unified, or in the way that James objects, by disputing the plausibility of the notion that consciousnesses and feelings can combine. If we are to take any of these seriously we would not expect facts about consciousness to be in any sense

dependent upon observations that rule out combining or dividing consciousnesses, because we would assume that such a thing could not happen in any case.

[3.3.3] A response to *a priori* objections

Even if we were to accept such *a priori* arguments wholesale, this does not suffice as a defence for consciousness science because *a priori* constraints on what must be possible in terms of the number of consciousnesses present in a situation are also constraints on what can possibly be discovered by empirical science.

In demonstration of a distinction between a conclusion justified by *a priori* reasoning and by empirical observation, there have been scientists who have posited that the split-brain is an example of consciousness switching between the two hemispheres (i.e. Levy 1977) and philosophers who have agreed (i.e. Bayne 2010). Levy's conclusions, it must be noted, arose from observations of certain brain activity only taking place in one hemisphere at a time. His model was thus available to empirical testing because there are specifiable observations that would demonstrate it to be false. Bayne, on the other hand, argued that the unity thesis is necessarily true, where a subject can only be in one complete conscious state rather than two divided ones (ibid.: 16). He argued for the truth of a tripartite conception of experience, which is of something being considered one unified experience if it occurs within the same subject, has the same content and occurs at the same time (ibid.: 24). Since Bayne regarded a subject as being a particular organism (ibid.: 9-10), he dismissed "two-streams" accounts of the split-brain, which regard consciousness as being divided when the corpus callosum is severed, primarily due to its incompatibility with the tripartite conception of experience (ibid.: 203). His main reason for dismissing the two-streams account then came not from observation or evidence but from its incompatibility with his conception of experience.

It is important to note that Bayne's philosophical support for such an interpretation of the split-brain, despite being the same as Levy's, was not support for an empirically confirmed conclusion. Levy's position was falsifiable whereas Bayne's was not. When presented with possible evidence of the presence of separate indicators of the presence of consciousness, Bayne countered it with *a priori* argumentation to the contrary. Although Bayne referred to empirical evidence to make plausible his conceptual framework regarding the unity of consciousness in ambiguous cases, he did not allow the empirical evidence to determine the truth or falsehood of his position.

For a science of consciousness to exist, there must be observable features that indicate the presence of consciousness. Regarding the potential persistence of any single consciousness, we have two options from here; either we can accept that using such indicators it may be possible to observe a number of consciousnesses constantly in flux, able to divide (as Sperry speculated was taking place in split-brain cases) and combine (as we might speculate would take place if we were to repair the corpus callosum of split-brain patients) or we refuse to accept that such a thing is possible.

Refusing to accept that such a thing is possible causes a problem because it entails that something observable like a behaviour or physiological state could not be an indicator of any particular consciousness, which results in intense difficulties. The same sort of thing does not seem to take place with regard to other empirically observable phenomena; we would generally assume that if you had an indication of the presence of a single body of a certain chemical at time A and then you had an indication of two separate bodies of that chemical at later time B, then provided no more of that chemical had been otherwise introduced we would assume the body to have divided into two. To relate these concerns to a science of consciousness and the specific *a priori* arguments mentioned, some explanation will be required.

I will not assert that there are examples of consciousness being observed to divide or combine, but, even if there are none, if it is *a priori* knowable that consciousnesses cannot do

so then there also can be no examples, regardless of what is observed, and it seems perfectly conceivable that there are situations where empirical observation would present us with cases that a science of consciousness could not explain. One such problematic case would be this: a consciousness scientist could observe a subject giving conflicting reports, with one hand writing down that it is aware of one set of phenomena and the other hand writing down that it is aware of another distinct set. It could be observed that two different systems in the body, which do not communicate with one another, are responsible for the production of both reports, and this would be such a problematic case. The question is how a science of consciousness should deal with such a case if it observed it; to those who accept *a priori* constraints on the possibility of a multiplicity of consciousnesses, we would have to deny consciousness to one or both "writers". We would then have to refuse to accept that writing is an indicator of the presence of consciousness, regardless of the eloquence and apparent self-awareness of the writer. The trouble is that the same argument can be made for *any* indicator of the presence of consciousness; even our capacity to speak could conceivably be observed to be so divided (although somebody born with two mouths might be required for such an observation). Without indicators, though, we cannot measure the presence and absence of consciousness.

For Bayne and Strawson's approaches to be compatible with the science of consciousness then, we would have to deny that such a case as the above could be observed, and yet without *a priori* knowledge of the specific structures that correlate with the presence of consciousness, which, as well as being highly implausible would erase any requirement for a science of consciousness in the first place, we have no means of ruling this out.

If instead we were to argue that Strawson was correct and that there is nothing to a consciousness aside from it being a state of a single body, and furthermore that we can empirically identify the presence of a single body, implying that even in the above example we could identify the presence of a consciousness with an empirical indicator, we would still be required to explain the apparent divided awareness. We would need to posit, for instance, that a single body has a single consciousness but that we may observe that single body (and

consciousness) having two distinct sets of conscious state simultaneously. Yet, describing the situation in this way does not help; if a consciousness can honestly report being unaware of a conscious state that it is nonetheless in then there is no telling how many unreported conscious states are present at any time. If this is so then we would still be unable to tell whether a conscious state is present or absent when a subject fails to report being in a conscious state because their report would not necessarily be indicative of the actual presence or absence of conscious states. A science of consciousness would thus still be impossible.

A different problem arises if we accept Kant's argument that consciousness must be unified, but it is also related to the presence of indicators of conscious states. Let us return to our previous problematic example and add some detail. We could observe a subject at time A giving a single, ordinary set of reports about its awareness. Then, at time B, we could observe the same subject giving two incompatible sets of reports, such as indicating with one hand that it is in a certain set of conscious states and indicating with the other that it is in a totally separate set of conscious states. When we ask the subject what they recall experiencing at time A, we might observe one hand reporting half of the memories the subject at time A had and the other reporting the other half of the subject's memories. If we were to observe this and accept Kant's reasoning regarding the unity of consciousness, we could not determine that our observations entailed that the individual's consciousness had divided into two separate consciousnesses. There are two ways that we could avoid this conclusion, neither of which will suffice to allow us to unproblematically continue with the science of consciousness that currently exists. We could first argue that there are not two separate consciousnesses at all, which is problematic for reasons already explained with regard to Strawson and Bayne's position.

The second way we could avoid the conclusion that a single indicator at time A becoming multiple indicators at time B does not entail a division of consciousnesses is if we assume that there are multiple consciousnesses but that only one of these consciousnesses is a continuation of that existing at time A, or else that neither of them are. So, rather than a

single consciousness becoming two consciousnesses, there could either be two wholly new consciousnesses that have come into existence at time B, or one new consciousness has come into existence at time B.

The problem with this case is that there would be no means for us to determine which of the multiple consciousnesses is the same as that existing at time A and which is numerically distinct. Wiggins outlined a hypothetical case where an individual's brain is cut into two hemispheres, both of which are then housed in two individual bodies, in order to put into question the notion of a person's identity continuing over time (Wiggins 1967: 50-58). If we imagine this as a more clear-cut case in which our possible observations of consciousness division could take place, the problem that arises when we consider consciousness as being necessarily singular and unified can be made clearer. Wiggins ruled out the possibility that they could both be identical to the person with both hemispheres who preceded them because, since identity is a transitive relation, this would entail that they should be identical to one another (ibid.: 53). As Parfit highlighted, none of the other possibilities are any more satisfactory (Parfit 1971: 5). If the person with both hemispheres is identical to one but not the other, there is seemingly nothing to determine which it should be identical to. Of most direct relevance here, there would be nothing observable that could answer this question for us; if both retained some memories and personality traits from the person who came before them but not others, we would have just as much claim to regard one set of memories traits as essential for the continuation of identity as the other. If the person with both hemispheres is non-identical to either of the two people resulting from his bisection, this entails that, even though a person can have one hemisphere severely damaged and still be the same person, the fact that *both* hemispheres have managed to continue to function strangely results in that person failing to survive. As Parfit asked, "How could a double success be a failure?" (ibid.: 5).

Parfit's suggested solution was to separate the notion of personal survival from one moment to another from the notion of identity, and so to state that, even though neither of the

two single-hemisphered people are identical to their double-hemisphered predecessor, the double-hemisphered person has *survived* in the two subsequent bodies (ibid.: 8-10). While identity is a one-one relation, survival can be a one-many relation (ibid.: 10).

It is not necessary though for us to accept Parfit's solution in order to see why Kant's reasoning would render a science of consciousness impossible. If we believed that consciousness was necessarily unified across a person's lifespan and further noticed indicators for the presence of consciousness in two individual bodies that seemed in some ways both physically and psychologically continuous with a previously singular consciousness, there is nothing by which we could determine whether the singular consciousness had ended or had continued in one of the two bodies. All we would know is that it could not have continued in both. From the fact that this could occur even within a singular body provided there were two sets of indicators for the presence of consciousness, neither of which appeared to be aware of the contents of the other, it would follow that we could not tell where one consciousness ends and an entirely new one begins. This means that we could not tell where a conscious state ends, and thus even if we accept the unity of consciousness endorsed by Kant we are still unable to determine what the empirically observable correlates of a conscious state are.

Finally, regarding the possibility of consciousnesses combining, this again produces a problem with regard to the possibility of a science of consciousness that relates specifically to our capability to find indicators for the presence of consciousness. To give a third variation on our hypothetical problematic case, we could observe a subject at time B writing two distinct sets of reports with either hand, describing an awareness of two distinct sets of conscious states. That subject could then, at time C, begin providing a singular set of reports that include the memories of all the states described by the individual hands at time B. In such a case, were we to assume that consciousnesses cannot combine, we would have to assume that at least one of the consciousnesses present at time B is no longer present at time C. This makes it impossible for a science of consciousness to measure whether even the consciousness they

are studying persists or rather ceases to exist since nothing about the subject's reports at time C indicate which of the two subjects at time B it is a continuation of, or indeed whether it is a new consciousness entirely. They would never be able to observe which conditions suffice for the presence of consciousness or the presence of a particular conscious state because they could not know if those consciousnesses or their conscious states had ceased entirely while they observed.

In brief, *a priori* constraints seem to rule out such speculative possibilities but only under the condition that these constraints eliminate the possibility that we can derive the number of consciousnesses present in a singular case from the number and character of indicators of the presence of consciousness. This means that the price we pay for adopting such constraints is the elimination of the possibility of the sole means a science of consciousness has of determining the presence of consciousness and conscious states through empirical observation.

Since, then, we have to allow for the possibility of observing consciousness division or combination in order to allow for a science of consciousness at all, this makes it conceivable that when somebody reports being in a certain conscious state and then reports no longer being in that state, their consciousness may have divided into one experiencing that state and one that is not. This coupled with the fact that we cannot rule out consciousnesses that fail to produce recognisable conscious behaviours means that we have no means to determine whether a consciousness has divided even if it has given no observable indication that it has. This would entail that if I want to claim that I am only studying your consciousness I would still be unable to do so since I could not tell whether your conscious state has ended even when you report that it has ceased. This is because I could not rule out that your consciousness has simply divided and that one of the new consciousnesses resulting from the divide, each of which is a continuation of the same consciousness pre-division, is in the conscious state that you are claiming has ceased.

[3.3.4] The plausibility of multiple consciousnesses

Even while we may have to admit the possibility that consciousnesses can divide and combine and that there may be multiple consciousnesses in an organism at any time, we may be inclined to brush off this possibility since we do not really know that such a thing happens anyway.

This sort of objection is the same sort considered about the possibility of consciousnesses that cannot produce recognisable conscious behaviour in section [3.2], and so the responses to this objection will run along the same lines. I will briefly give these responses here.

There are no grounds provided by verificationist reasoning to object to the possibility that a single consciousness cannot divide into one consciousness producing recognisable conscious behaviour and another that fails to do so. Such a thing cannot be observed, but the contrary claim that consciousnesses cannot or do not do so equally cannot be supported by empirical evidence. The fact that a science of conscious states depends crucially upon it being false that a single consciousness sometimes so divides shows that this science is unusually dependent for its truth on the falsehood of speculative possibilities.

The possibility of multiple consciousnesses equally cannot be dismissed as improbable because, without first ascertaining what the empirical correlates of consciousness division are, we cannot know how frequently such a thing occurs. It could be as rare as to require the division of hemispheres or as common as to only require a small cluster of neurons to function independently; without any means of gathering empirical evidence on the subject, the likelihood is unknowable.

Since a science of conscious states has no means of dismissing the possibility of consciousness divisions, and crucially depends upon it being false that such a thing takes place, this means that it has no means to support the claims it does make.

[3.4] We cannot identify empirically observable correlates of PC states.

The above considerations have largely focused on consciousness in general, so it is worth bringing this back to discussion of PC states, since these are our focus in evaluating the plausibility of a science of consciousness. As argued in this chapter, where a PC state is present, there is no way to determine through empirical observation where the state by which our conscious state can be communicated is distinct to the state by virtue of which our claims to be in a conscious state are true. If the two are distinct, there is no possibility of identifying the empirically observable correlates of this conscious state; we will never know which aspect of the overall physiological or cognitive state corresponds to the conscious state because that same conscious state could persist in very different circumstances and we would not know. Even this presupposes that there is a clear criterion by which we can determine if a state is conscious – the capability to produce recognisable conscious behaviour – and yet, as argued in the previous chapter, this criterion is so hopelessly vague that it risks making the idea of a PC state just as vague.

The same problems apply to a wide variety of approaches to such a correlative approach. Two more recently defended scientific theories of consciousness are information integration theory (IIT) and semantic pointer competition (SPC), both of which can be shown to fail to avoid the criticisms given above.

IIT is a theory that results from the comparison of PC states with non-PC states, such as a human receiving a blow to the head, being administered general anaesthetic or nonhuman states such as those of a camera (Tononi 2004: 42). The difference between the former kind of states and the latter is that information is integrated in paradigmatic states of consciousness to a much greater degree, meaning that the elements in any of the latter states work largely independently; you could cut a sensor chip in a camera into its individual photodiodes without

this affecting the performance of the camera, but separating the processes of your brain involved in paradigmatic states of consciousness would have “disastrous effects” (ibid.: 42). The combination of this description of functionality and descriptions of our phenomenology as being of elements that cannot be experienced independently, such as both halves of our visual field, and shapes and their colours, is stipulated to give us reason to believe that the two rely on linked processes. As such, the amount of consciousness in a given system is a product of the level of information integration in that system.

Now, the main bulk of my criticisms regarding the above scientific theories have centred on the fact that they rely on contrasting conscious and non-conscious processes, which IIT in a sense denies. A great number of physical processes, even relatively simple ones such as those responsible for temperature readings on a thermometer, involve some level of information integration and are thus, according to IIT, also conscious processes. As such, we might determine that IIT does not rely on the contrast I have criticised above, given that many of the processes contrasted in IIT will be taken to also be conscious.

However, it is clear that Tononi’s position was developed under the assumption that there is a contrast between the two kinds of process. Indeed, his entire justification for determining that information integration and consciousness are related was that there is a distinction in terms of the presence of consciousness between states such as a human in an ordinary state of wakefulness and those of a camera detecting the presence of light. Although his position may have gone on to then state that there is also consciousness in the latter case, it could only do so once this initial stipulation had been accepted. This appears additionally problematic because, once we accept that both states are conscious, it appears we no longer have justification to believe that information integration was important at all.

Tononi’s means of avoiding this was to stipulate that, although there is consciousness in all cases where information is integrated, there are different levels of consciousness such that a system where information is more integrated is more conscious. This position suffers from the difficulty that it must be assumed sense can be made of the idea that one

consciousness can have a lower level of consciousness than another. This is not something I am going to attempt to make sense of here. Even if we can unproblematically state that there can be different levels of consciousness, this defence simply results in the same problems with previous positions. Tononi regarded himself as comparing more conscious processes with less conscious processes, relying crucially on the assumption that it must be processes with higher levels of consciousness that produce recognisable conscious behaviour. Without this assumption, it would have been impossible in the first place to identify information integration as related in any important sense to consciousness and thus the theory would be without empirical justification. For instance, Tononi stated that a theory of consciousness needs to explain why changes in thalamocortical regions are so important for conscious experience, "whereas changes in neural activity in cerebellar circuits are not, given that the number of neurons in the two structures is comparable" (Tononi 2005: 109). Such a thing only needs to be explained provided we assume that such a thing has been observed to be the case, and it has already been discussed in some depth why the importance of thalamocortical activity in conscious experience has no justifiable scientific basis. Describing the same situation as one of comparing greater or lesser quantities of consciousness does nothing to avail these problems; if we have a problem finding indicators for the presence of consciousness then this problem applies equally to finding indicators for the presence of specific levels of consciousness.

SPC has pitted itself against IIT on the grounds that stipulated conscious states existing in entities that produce no recognisable symptoms of the presence of consciousness are empirically problematic (Thagard & Stewart 2014: 75). Of course, this justification runs into the problems already addressed in [3.2] – [3.3] regarding conscious states failing to produce recognisable conscious behaviour, verificationism, and the likelihood of any state that fails to produce such effects being conscious. SPC has then attempted to justify its position by reference to the phenomena that it means to explain, including the cessation of consciousness in sleep, anaesthesia, concussion, strokes and seizures (ibid.: 78), clearly running into the

problems identified in [3.2] with determining whether a state is nonconscious. As such, the conclusion that they arrive at, that semantic pointers, "which are representations that can function as symbols while retaining connections to sensory and motor representations" (ibid.: 74), are essential for consciousness, given that this is posited specifically with the above advantages, has no supporting justification.

The above criticisms, although applicable to existing scientific practice, are equally devastating for any future of such practice. Since there must be some means for this science to identify the presence of PC states, the sorts of arguments I have offered here will present a challenge. Unfortunately, even if these criteria are not necessarily as vague as the capability to produce recognisable conscious behaviour, it will always be possible that consciousness could exist in the absence of whichever empirical indicator is adopted, and the only way to deny this is to accept an *a priori* argument that renders an empirical study of consciousness impossible.

[3.5] Summary

We have arrived at the conclusion that the science of consciousness is unable to empirically distinguish between the presence and absence of PC states. The ambiguity of the very idea of a PC state has been revealed; for a conscious state to be paradigmatic, it must be the sort of state that could communicate its presence, but this tells us nothing about whether the state that is being indicated would be present or absent where that capacity to communicate is not present [3.1].

We explored the possibility that consciousnesses may sometimes not be able to produce behavioural indicators of their presence [3.2]. It was showed how this would render a science of consciousness impossible in [3.2.1] and showed that we cannot simply deny the

possibility that there could be such cases without similarly rendering a science of consciousness impossible in section [3.2.2].

Similarly, we saw how the presence and absence of conscious states cannot be ascertained by the presence or absence of behavioural indicators [3.3]. We explored the possibility of consciousnesses dividing and combining to make it impossible to ascertain where a conscious state ends and where it begins [3.3.1], and we found in sections that we cannot easily deny the possibility that such cases could occur without similarly rendering a science of consciousness impossible [3.3.2] – [3.3.4].

We concluded finally that a science of consciousness cannot contrast the presence and absence of consciousnesses or conscious states and so cannot complete the task described in [2.3] of identifying empirically observable correlates of PC states [3.4].

In the next section, we will besiege the last bastion of the science of consciousness, which is that it should at least be able to give us the timing of conscious states so that we know that a certain conscious state occurs at the same time as a certain neurophysiological or cognitive state. Even this modest goal, we shall now see, is unattainable.

4

Timing Conscious States

[4.1] Libet et al.'s subjective delay

We might consider it to be the case that whether or not we can determine which empirically observable events correspond with which conscious states, we can at least determine that the PC state a subject is in at a particular moment can be found to be simultaneous with certain other empirically observable events. To see just how problematic this is, I will use a famous experiment on the timing of conscious experience as an example. In this experiment, subjects had electrodes placed on their head, arm and hand (Libet, Wright & Gleason 1982). They were asked to watch a spot of light revolving in a clockwise circle and to flex their fingers or wrist whenever they felt like doing so. They were also asked to report the position of the light when they became aware of the urge to perform the voluntary action. The difference between the subjects' reported time and the time the electrodes measured an increase in activity known as a "readiness potential" was measured (ibid.: 323–325).

The time the subjects reported becoming aware of the stimulus (as indicated by the position of the rotating light) was roughly 350ms after the readiness potential was measured. These findings were repeated in later experiments (Libet et al. 1983). The researchers concluded that "cerebral initiation of a spontaneous, freely voluntary act can begin unconsciously, that is, before there is any (at least recallable) subjective awareness that a 'decision' to act has already been initiated cerebrally" (ibid.: 106).

The conclusions drawn from these studies by the researchers conducting them and others have been challenged in several ways, but many, perhaps most, have focussed on the implication that these experiments disprove the existence of free will. It has been argued that

the idea that there is a particular conscious event a subject can identify as the initiation of the intention to act is debatable, and that subjects of such experiments may have to invent such a conscious event, such as mentally saying the word “now!” in order to have something to correspond to the “clock” (Mele 2009: 34-35). The assumption that free will requires such specific conscious events seems to be questionable (ibid.: 36).

It has been convincingly argued that these experiments only seem to disprove free will if you ignore fact that the reasons for the individual’s actions were developed over a longer period of time, which involved the decisions to take part in the experiment and to obey the experimenter’s instructions, which in turn involved a continued conscious awareness to act in this way over a longer period of time, and was presumably a significant factor in the decision to make the spur-of-the-moment decision to flick their wrist at a certain time (Tallis 2011: 247–250; Satel & Lilienfield 2013: 133–137). It is thus only by taking the action out of the larger context within which the action occurred that we can arrive at the conclusion that it is only the immediate causal antecedents that are relevant to the decision for a participant to flick their wrist.

The criticism that I wish to focus on here though, being directly relevant to our current task, is that the claim that, even if there is a conscious event we can specify as the point where we decide to act, the notion that we can determine a specific time that that conscious experience took place relies on some very questionable assumptions. As Dennett observed, Libet’s experimental method did not take into consideration that there could be a difference between the time it takes for you to become aware of the time shown on the clock and for you to become aware of your decision to act (Dennett 2003: 232-234). That is, it assumed that you have a centre of consciousness that receives visual information more or less instantaneously but where the timing of its receipt of other information, such as that associated with decision-making, is up for question. Dennett has argued against the notion of a centre of consciousness before, referring to it as “the Cartesian Theater” (Dennett 1991a: 101-138), but in this case he agreed to take the idea seriously for the purposes of his criticism (Dennett 2003:

232). Essentially, Libet's conclusion depended on his subjects having a centre of consciousness in a specific location in the brain. If it is located close to the visual centre, we would expect the time that we subjectively report being aware of the clock to be comparatively close to the time actually shown on the clock, and thus for the time the subjects report being aware of their conscious decision to act could be accurate. If, however, it is located closer to where decision-making takes place in the brain, we would expect the time to be further off. If your centre of consciousness does not inhabit a single location and moves around in your brain, this would entail that the delay from either area could be larger or smaller. Otherwise, your centre of consciousness could be in a different location to either the decision-making processing area or the visual processing area, in which case information from both could be expected to be delayed by perhaps different amounts of time. Dennett presented these and other possibilities (ibid.: 232–237), and many of the above cases would mean that an estimate of when your conscious experience of decision-making took place would be quite different from the time estimated based on the visual information you received about the clock.

Libet responded to these criticisms based on the possibility of alternative mistimings such as sensory delays by stating that commentators appeared to have missed a "critical control feature" in the experiments (Libet 2003: 326), which was that subjects were asked to judge when they felt a sensation on their hand following electrical stimulation to their hand, and their reports were found to be misjudged from the actual stimulation by only 50ms (ibid.: 326). The problem with this response and the description of this condition as a "control" is that it does nothing to allay any concerns that subjects cannot accurately report the time at which their conscious states occur. Rather, it simply tells us that subjects can (almost) accurately report the time at which stimuli are present. Yet, a subject could presumably accurately report when their skin was stimulated even if their conscious state was delayed by longer than 50ms if the subject was unaware of this delay.

Another interesting interpretation of the experimental findings was suggested, following the result that subjects only report conscious awareness after 500ms of cortical stimulation,

but they report this stimulation occurring simultaneously with the time the stimulation commenced. The experimenters suggested that this evidence was of antedating of subjective timing, or "backward referral" (Libet et al. 1979). The reasoning behind this seems to be that conscious awareness does not come into existence without 500ms of cortical stimulation, which is a highly dubious assumption as illustrated here in earlier discussion of identifying the distinction between the presence and absence of consciousness, and that the subject reports the timing commencing prior to the point the sufficient neural conditions have been reached, which is also unclear as illustrated with regard to Dennett's alternative hypotheses above. The authors do not seem to have been suggesting that the subject merely *judges* the conscious state to have taken place earlier than the sufficient neural state has taken place, but rather that the subject is correct in their judgement; as stated in the summary of the aforementioned paper:

> But a dissociation between the timings of the corresponding 'mental' and 'physical' events would seem to raise serious though not insurmountable difficulties for the more special theory of psychoneural identity. (ibid.:222)

If the judgement of subjective timing were due to anything other than a conscious state taking place prior to the sufficient neural conditions for its presence, then it is not clear that this particular finding should present any difficulty for the identity theory, suggesting that the experimenters were certainly taking seriously the notion that conscious states could avoid following the temporal structure of neural events. Indeed, if such a thing were to be taken seriously, the threat would not simply be for the identity theory, but for any theory that requires conscious states to be realised by neural states, such as a functionalist view that conscious states are identical with a certain role performed by any appropriate state, of which neural states would be an example.

Rather than rejecting the notion of antedating as being unnecessarily exotic, it may be possible for us to take seriously the idea of subjective antedating within an empirical

framework, as Penrose attempted to do. Penrose attempted to describe how consciousness could be related to brain activity while simultaneously able to follow a different temporal ordering to that of the measured neural events (Penrose 1989: 446). The mechanism he suggested is responsible for this is Correct Quantum Gravity, a "sought-for theory" that unites quantum physics and general relativity (ibid.: 348), which may be able to describe how conscious experience may constitute "some kind of actual contact with Plato's world of ideal mathematical concepts" (ibid.: 446). Given that "Plato's world is itself timeless," this suggests that a full understanding of consciousness cannot be grasped by simply placing conscious states in a temporal structure (ibid.: 446). His inferences seem to be derived from the following considerations:

- We have access to mathematical truths that can be gleaned non-algorithmically, such as propositions demonstrated using Gödel's theorem (ibid.: 118, 416).
- Non-algorithmic mathematical truths could be made sense of given a form of Platonic realism (ibid.: 94-98).
- Neural computations are usually posited as being algorithmic in nature (ibid.: 392-396).
- Neural signals could potentially be triggered by quantum events (ibid.: 400-401).
- Correct Quantum Gravity must not be describable in ordinary space-time terms (ibid.: 447).

These joint considerations would place a potential formulation of Correct Quantum Gravity as a viable candidate to describe how we can arrive at non-algorithmic mathematical truths, and how our judgements about mathematics can be determined by existences in a world of mathematical forms. Presumably, the effects of the mechanism described in a theory of Correct Quantum Gravity must be able to affect neural processes in a nonalgorithmic way such that the judgements that we form are affected by truths about these forms.

There are several problems with Penrose's position. One is that it not only relies on a formulation of Correct Quantum Gravity that has not been given, but it gives additional work to this undiscovered formulation by supposing that it should also be capable of solving problems related to mentality. Additionally, Penrose's position relies upon the truth of a form of Platonic realism without this following from any engagement with the philosophical difficulties of this position. Furthermore, it posits an unproven mechanism to explain how subjective antedating could take place when it seems that we could make sense of the data without relying on such a thing, such as by using one of Dennett's alternative suggestions as to how we might have arrived at our judgement of the temporal sequence of events. These considerations suggest that Penrose's primary objective was not to find the simplest or most plausible explanation of the data from Libet et al.'s experiments but rather to explain how we can think non-algorithmically.

We do not need to posit amendments to our understanding of physical reality to maintain a connection between conscious states and empirically observable states while rejecting interpretations of Libet et al.'s experiments that require antedating of conscious states. The problem is that rejecting antedating without justifying this using some such amendments requires us instead to make *a priori* assumptions that would render it impossible to empirically determine the temporal ordering of conscious states.

This problem arises when we consider how it is that we can determine that one conscious state occurs at the same time as another empirically observable state. If I imagine the first-person perspective of a subject in one of Libet et al.'s experiments, I might be having a particular set of conscious experiences that seem to follow some ordering, such as the experience of perceiving a clock showing various times, and then that clock showing a certain time alongside the urge to flick my wrist. How, then, am I to determine whether my experiences have happened at the same time as a given event existing independently of my consciousness that is supposed to have caused my experience? I cannot experience my conscious state and

the state that causes it separately in order to observe the time difference; indeed, this would need to happen *within* my experience anyway.

The same applies to the experimenters observing me. They can determine when certain neural events occur, when I flick my wrist, and the time I report. What they cannot determine is when my conscious state occurred in relation to these events; they equally cannot compare my conscious state to observations of my brain state, behaviour and surroundings and they only have my reported time as a basis for making any judgement about my conscious state.

The idea is contentious that we can compare two things that I will refer to as “inside” and “outside” consciousness. Things lying inside consciousness are those that form part of our subjective awareness, and things lying outside consciousness are those that exist independently of consciousness but that may be causally relevant to our conscious states. The denial that these two things can be referred to separately leads to the view that experience is transparent, which will be discussed shortly. We can adopt different descriptions for things lying inside and things lying outside consciousness, but this does not solve the problem either as we will also discuss shortly.

In all, I will consider four major responses we could give to the view that the timing of conscious states cannot be compared with that of non-conscious states, all of which are inadequate. They are:

1) Conscious states are transparent.
2) My conscious state of perceiving X is caused by the presence of X.
3) Our concept of time is derived from the succession of conscious events.
4) There is a structural resemblance between conscious events and external events.

These responses, particularly 2–4 all rely upon the causal theory of perception, as defended by philosophers such as Locke and Russell, and as such discussion of these two philosophers will form part of the upcoming discussion.

[4.2] Transparency

The problem stated thus far might seem to make some presuppositions about the nature of consciousness, particularly pertaining to how we become aware of conscious states. To state that we cannot compare the states of the world around us with states of consciousness seems to presuppose that there are two sets of states – those of the world and those of our consciousness. We could attempt to avoid this by arguing for the transparency of conscious states:

Transparency

> Nothing is available to introspection other than the objects represented as in one's environment, and the properties they are represented as having. (Speaks 2009: 542)

Transparency is opposed to the view that the states we become aware of through introspection are distinct to the states that form the content of our perceptual states, which I will refer to as opacity and will discuss further in [4.3].

In his argument against idealism, Moore challenged the view that what we perceive in experience is distinct from the properties of the objects external to the subject. He stated that common arguments for idealism depend upon the joint contradictory suppositions that experiences are distinct from other non-experiential things and that it is impossible to

distinguish an experience of a property and a property existing outside of experience (Moore 1903: 442). He argued that the confusion arises because when we attempt to refer to a sensation, such as “the sensation of blue,” we seem to “look through it and see nothing but blue” (ibid.: 446). As such, the mental fact seems to be, in a metaphorical sense, “transparent” (ibid.: 446).

The position has subsequently been defended as forming the basis for representationalism, as is plain from the above quote by Speaks. Representationalism is the view that the character of a conscious state is determined by the content of a particular representational state. Tye, a prominent defender of representationalism, argued for transparency in the following way:

> If you are attending to how things look to you, as opposed to how they are independently of how they look, you are bringing to bear your faculty of introspection. But in so doing, you are not aware of any inner object or thing or event. When you introspect your visual experience, the only particulars of which you are aware are the external ones making up the scene before your eyes. You are not aware of those objects and a further inner object or episode. Your awareness is of the external surfaces and how they appear. (Tye 2002: 139)

If we accept transparency then it seems as though my above problem is misstated. Although Tye agreed that there is no separate means by virtue of which I can become aware of what lies inside and outside of my awareness, if conscious experience is transparent then the states we are aware of through introspection are the same as those we are aware of through perception, in which case there is no need for the two to be independently examined and correlated.

What then would we say in response to Libet et al.'s claim to be able to locate the time at which a conscious state has occurred? The problem would remain since the conscious state would still be a state of the subject and so we would have to determine the timing of the conscious event to ascertain whether it is simultaneous with a particular nonconscious event. It would be to misstate representationalism as claiming that the character of the conscious state of perceiving a tree is dependent upon the tree itself; the conscious state is identical to the representational state and as such it is an aspect of our psychology, which we might take to be a certain cognitive state or neurophysiological state, that determines the character of a conscious state. For instance, in attempting to determine which entities are conscious and which are not, Tye stated, "a mental state is phenomenally conscious just in case it has a PANIC–a Poised, Abstract, Nonconceptual, Intentional, Content" (Tye 2000: 172). As such, conscious states do not belong to external events under representationalism any more than they do under a view that considers experience not to be transparent; rather, they depend upon being *about* certain things that may be external to the entity in the representational state. Conscious states are still, under such a representationalist view, realised by entities that represent things in the correct way. Tye ruled out consciousness in a Venus fly trap and caterpillar because they do not meet the stated criteria (ibid.: 172-174), and so if a caterpillar was to find itself in a physically identical environment to one that you or I might find ourselves in, the conscious states we are in when we perceive our environment would be absent. As such, a conscious state is a state of the entity representing its environment and as such the question can still be raised as to whether the conscious state occurs at the same time as an event occurring outside of consciousness.

There is, then, no reason to suspect that transparency offers us any means of avoiding the stated problem. If I were to represent a state as occurring at a certain time two hours before it occurred, or even two hours afterwards, representationalism would not give us a means of determining that this had occurred merely by examining my own representational state. Indeed, whether through perception or introspection, all I would be able to tell is that the

content of the representational state is an event occurring at a particular time, not when the representational state itself occurred.

[4.3] Conscious states of perceiving X being caused by the presence of X

For us to interpret Libet et al.'s experiments as giving us evidence that our conscious states occur at the time the experimenters suggest, we would have to assume that there is a specific conscious state that we would describe as a state of perceiving X that always follows from a subject bearing a certain relation to the X being perceived, and that this relation is possible to identify through observation. Their conclusion would not follow unless, amongst other things, we assumed that:

a) The subjects are in the conscious state of visually perceiving a clock.

b) The presence of that clock caused that conscious state.

Ayer used an example given by Grice, where a causal theory of perception would state that an individual perceiving a hand does so because a hand is causally producing this perception, "But the trouble with this is that such an example only works if it is already established that I am looking at a hand" (Ayer 1977: 113). That is, we cannot introduce the concept of a hand purely on the basis of my sensory experience because we would have to know already that perception of an object requires the presence of that object. To add a further point, there are quite a few assumptions that need to be made.

We must assume that there is some sort of relationship between our experience and the object producing our experience such that the experience is "of" that object. To see why this is the case, let us imagine that the conscious state one subject has when looking at the clock is the same conscious state another has when feeling a sharp sense of pain. This latter subject then described the clock showing a certain time, but their conscious state was the same as the former subject's experience of reporting being in pain. In such a case, the

conscious states of each individual may result in certain behaviours, but there would be such a discrepancy between what the subject was experiencing and what their reports suggested they were experiencing that there would be no way for either the experimenter or the subject to correlate the two. Indeed, the experimenter would be unable to tell if the subject was even having the same experience when they reported the same thing, or if the subject was sometimes conscious and sometimes unconscious, and so on. The subject would be unable to tell what was actually happening in reality around them and would be responding to their own conscious states, which would be seemingly unrelated, like somebody believing they are typing an essay but are actually piloting an aeroplane. Since this would clearly make it impossible to correlate conscious states with non-conscious states due to the seemingly unrelated content, we clearly must assume this is not the case. The question is whether we can make this assumption based on empirical observation or if it requires a purely *a priori* defence.

Ayer seemed to adopt a transparency viewpoint when suggesting that the objects within our perception are made up of the particles posited by science (ibid.: 123), but there are plenty of examples of non-transparency, or opacity, views of experience through the history of philosophy.

Opacity

> Nothing is available to introspection or perception that lies outside of our conscious experience.

Prominent examples of opacity theories can be found in the work of Locke and certain works of Russell.

Locke argued that the qualities of things in the world can be distinguished from the ideas of those things we have in our minds, and by "ideas" Locke means "the immediate object of perception, thought, or understanding" (Locke 1689: II.VIII.8). He wrote:

> To discover the nature of our ideas the better, and to discourse of them intelligibly, it will be convenient to distinguish them as they are ideas or perception in our minds, and as they are modifications of matter in the bodies that cause such perceptions in us: that so we may not think (as perhaps usually is done) that they are exactly the images and resemblances of something inherent in the subject (ibid.: II.VIII.7)

It is by virtue of this distinction that Locke further distinguished between primary and secondary qualities, which are, respectively, the qualities that we must attribute to matter whether or not that matter could be perceived (ibid.: II.VIII.9) and the qualities that we must attribute to matter by virtue of which matter can produce sensations in us (ibid.: II.VIII.10). As such, many view Locke as believing that what we perceive are ideas that are distinct from the objects that cause them in us, such as Bennett, who stated that Locke's view places the objective world "beyond our reach on the other side of the veil of perception" (Bennett 1971: 68).

This seems a reasonable, if not universally agreed, reading of Locke (with a direct realist alternative argued for in Rogers (2004) for instance). In any case it is this view often attributed to Locke that qualifies as an opacity view of consciousness. According to this view, that which we are aware of through our conscious experience is not states belonging to the world that cause our experience but rather states belonging to ourselves.

Similarly, Russell argued that what we are aware of in experience, even in cases of veridical perception, cannot be objects in the external world that cause our experiences, but rather:

> If we define a piece of matter as a set of events, the sensation of seeing a star will be one of the events which are the brain of the percipient at the time of the perception. Thus every event that I experience will be one of the events which constitute some part of my body. The space of (say) my visual perceptions is only correlated with physical space, more or less approximately; from the physical point of view, whatever I see is inside my head. I do not see physical objects; I see effects which they produce in the region where my brain is. (Russell 1914: 99)

While Russell believed that sensations of perceiving colour are identical with the actual existence of colours, forming part of the physical world "and part of what physics is concerned with" (Russell 1921: 141–142), this is not a transparency view because Russell's defence of the causal theory of perception led him to conclude that "what the physiologist sees when he examines a brain is in the physiologist, not in the brain he is examining" (Russell 1927: 320). As such, while he believed that our perceptual sensations form a part of the subject matter of physics, this would perhaps be better seen in the same sort of way reductive physicalists understood the two to be continuous; they are both physical and our sensations are, specifically, neurophysiological states. The qualities of our perception, then, are qualities of our brain states, and the inferences we make about the world outside of our brains can only be made on our examination of structural properties of those immediately apprehended qualities (Russell 1921: 253).

Now, the problem with opacity in relation to the current issue is even worse than that of transparency. If we are only ever aware of events in consciousness, then we have no means by which to time when these events occur in relation to events lying outside consciousness. Indeed, if entities in the world are causing the events lying inside consciousness, then we have no independent means of determining that the events lying inside consciousness are even

about those events and are not, for instance, being induced in us by a scientist who has our brains in a vat.

Russell's perspective is not so easily subject to such criticism because he argued that our knowledge of the physical world is derived from sense-data, which we will address in point (3), but if we were to take his view as advocating a form of opacity, which on occasion it certainly seems to be, we would be unable to determine that when we perceive a clock it is the presence of a clock that is causing that perception.

Most importantly of all, none of the above can be resolved via empirical means. That is, we are required to have an *a priori* refutation of sceptical hypotheses and an *a priori* endorsement of the perspective that what we are aware of in perception is that which has caused our perception. This cannot, of course, be demonstrated empirically because to do so there would have to be some observation we could make that would tell us that our conscious state is being caused by the particular entity we are perceiving rather than being something else, but *ex hypothesi*, we could only be aware of it being caused by something else if we were not in the same conscious state.

What this means is that we have to accept the truth of a non-verifiable proposition in order to claim that we can determine when conscious events occur in relation to nonconscious events, which is the following:

My conscious state of perceiving X is caused by the presence of X.

We have no way of determining merely from what we observe whether the above statement is true or false given that all our observations are compatible with the rejection of the above claim and the acceptance of a sceptical hypothesis such as brain-in-a-vat-type thought experiments. The problem with deferring to non-verifiable claims has already been discussed with relation to the claim that "All conscious states must produce recognisable conscious behaviour," and I do not wish to reiterate these criticisms here.

Rather, we must assume that there *a priori* reasons for its truth. The trouble with this claim is that (as has also already been discussed with relation to *a priori* reasons for rejecting

the possibility of a multiplicity of consciousnesses), deferring to *a priori* reasoning to eliminate possibilities equally removes the possibility of there being an empirical indicator that would be able to tell us such things. For instance, Libet et al.'s experiments require there to be some empirical indicator for the timing of a certain conscious state, which they take to be the subject's reported awareness of the position of a clock. If we assume that it is *a priori* true that my conscious state of perceiving a clock must lie between the presence of the clock and the production of my report, then we must subsequently deny that any observation could show this to be otherwise.

Let us imagine a rather baffling instance, for instance, that we were to observe an individual reporting a time on the clock face that was only present before the individual looked at the clock. If we were, then, to take the subject's reported timing of the clock face as an indicator of the timing of their conscious state, we would have to assume that the conscious state occurred prior to the subject looking at the clock.

Such an observation would show many known scientific facts to be incorrect and thus we have good empirical reason to believe that we will not make such an observation. For one thing, it would need to be accepted that it is possible for an individual to become aware of something through a mechanism other than the known senses. However, there is a world of difference between the claim that such an observation would run contrary to known physical laws and the claim that such an observation is in principle impossible to make. The fact that such an observation would run counter to known physical laws can be taken as a specification of what would be required to demonstrate such laws to be false; indeed, physical laws are useful in science largely because of the potential they give us to specify what an observer should observe in particular circumstances. I do not claim that the scientific community would abandon their accepted theories if presented with such an observation; as pointed out by numerous philosophers, there are plenty of examples of science observing such "anomalous" results and maintaining their original theories with perhaps *ad hoc* amendments (Kuhn 1962: 77–91, Lakatos 1970, Feyerabend 1975). I simply claim that such an observation would be

consistent with the abandonment of known scientific theories and would not present any incompatibility with the possibility of a science of consciousness in particular.

Such an observation would also run against our *a priori* assumptions about what causes our conscious states, and yet the situation here is rather different. The abandonment of such *a priori* assumptions in the face of such an observation would render a science of consciousness impossible. Thus we must either assume such an observation to be impossible, and thus assume that we know what we can observe prior to empirical observation, eliminating any use for a science of consciousness, or we must assume such an observation to fail to show that conscious states do occur at the time we originally supposed them to, meaning that reports of clock positions are not indicators of the timing of conscious states. The trouble is that no matter which empirical indicator we use, we would have to make the same assumption to maintain that a science of consciousness can tell us anything about the timing of conscious events. As such, we will always be forced to assume that no empirical observation could *in principle* tell us when conscious states occur.

It is worth stating with regard to this what was stated with regard to the possibility of unrecognisable conscious behaviour; this is not a case of me giving unreasonable demands to a science of consciousness such that it should be able to deal with speculative possibilities that no other science is required to. This is a case of a science that is unusually susceptible to thought experiments such that it must be able to demonstrate their conclusions to be false before it can assert that its own claims are true. I do not need to know that I am not a brain in a vat to believe in the truth of claims made by physicists; these are claims about the relation between observable events, whether or not these observable events are real or imagined. Indeed, if I was to discover that I am a brain in a vat, the relations between the observed events described by such sciences would still hold; they would simply be about a world that exists only in my mind.

Yet, I *do* need to know that I am not a brain in a vat if I want to assert that a conscious event occurred at a particular time in relation to the event that caused it. It is for this reason

that consciousness science is uniquely fragile; it requires us to hold a set of propositions that are impossible to demonstrate through science and that no other science must hold, otherwise its justification for adopting the claims that it does falls away and our claim to know that which the science tells us subsequently shatters.

[4.4] Deriving our concept of time from the succession of conscious events.

Another line of defence for consciousness science is to argue that we derive our concept of time from the succession of conscious events, and thus it is a misstatement to claim that a conscious state could occur at a different time to that which we think it does. I think this would be even more problematic for a consciousness scientist to maintain since it would plausibly mean that conscious states cannot be temporally correlated with other states at all. But there is a more decisive argument that can be used against this claim.

To lay out the claim properly, I will present positions by Locke and Russell again, followed by a brief note regarding Kant's position on time.

Locke suggested that the reason we have a concept of time at all is that we experience conscious states occurring one after another:

> I think it is plain, that from those two fountains of all knowledge before mentioned, viz. reflection and sensation, we got the ideas of duration, and the measures of it. For, first by observing what passes in our minds, how our ideas there in train constantly some vanish and others begin to appear, we come by the idea of succession. Secondly, by observing a distance in the parts of this succession, we get the idea of duration. (Locke 1689: II, XIV, 31)

Although Locke's goal was to demonstrate that we get all of our ideas from sensation or reflection (ibid.: II, I, 2), which may seem aligned with the consciousness scientist's commitment to empirical observation, the Lockean conception of time actually fails to support the view that a science can correlate our conscious states with other empirically observable states.

Russell also presented a conception whereby we derive the concept of time from relations between states lying within our perception, such as the relation between one sensation that has partially faded and another that has not, or "between a percept and a recollection, both of which occur at the same time" (Russell 1921: 254). Additionally, we derive our sense of time from the fact that we recollect processes in a certain order. (ibid.: 254).

There are two problems with such accounts of us deriving time from our own experience.

Firstly, even if my conception of objective time is derived from my experience of sensations passing in my mind, it does not follow that the two are identical measures of the passage of time. I might start out understanding numbers by counting my fingers, but it is quite clear that by the time I am able to count unaided I am no longer thinking of fingers. Indeed, I can come to understand new mathematical truths that could not be derived from finger-counting, such as an understanding of negative and imaginary numbers. We do not have to reject Berkeley's argument that we can never abstract away from experience altogether (Berkeley 1710: VI-XXI) to argue that we can use plainly apprehended properties in order to create a conceptual framework by virtue of which we may come to new understanding. Berkeley argued that time is nothing "abstracted from the Succession of Ideas in our Minds" (ibid.: XCVIII), constituting part of his argument for the view that our perception of reality is wholly produced and sustained by God (ibid.: LXVIII). Yet even in Berkeley's subjective idealism there is room for differing conceptions of time; if ideas are induced in us by God and our conception of time is produced by these ideas, it follows that the particular ordering of events we perceive is only that which God has produced in us and not necessarily that God

caused those ideas in that same order. If God caused us to experience one day's worth of events, there is no reason to suppose that it took God one day to cause those events to occur. As such, there is still a legitimate question as to whether the ordering of events outside of our experience is the same as that of events inside it.

In other words, even if our objective conception of time does follow from a certain succession of conscious states, it does not follow that we must consider these two conceptions to be identical. Thus, it is reasonable to ask how we can know that the two do correspond.

The second problem is that, if we accept such *a priori* argument for the two conceptions of time being linked, we would have to abandon any empirical indicators for conscious states. Either we determine that a conscious state occurs at a certain time by observing the effects of that conscious state at a similar time, or we assume that it must have occurred at the same time as certain external events based on *a priori* reasoning. So if the temporal order of a subject's conscious states were out of line with that of objective timing, such that a subject experienced what they believed to be a clock showing a certain time at *t*, but then their experience of the clock ticking a single second actually occurred at *t minus six hours*, we would either be able to observe this discrepancy through the presence of empirical indicators or else we would have to assume it through *a priori* reasoning. If the former, then we would have to reject the *a priori* argument for the simultaneity of timings. Without such *a priori* argument though we have no way to empirically determine which conscious states are evidenced by the presence of which indicators, and no empirical evidence can tell us this given that we cannot independently observe conscious states and their causes. If we accept the *a priori* reasoning though we equally cannot empirically determine which conscious states produce which indicators because such a thing would have to be known prior to any empirical investigation. Either way, a science of consciousness cannot tell us the timing of conscious states.

An alternative position placing time as dependent upon consciousness is the Kantian position. Kant argued that time "is not an empirical concept" since "neither co-existence nor succession would be perceived by us, if the representation of time did not exist as a foundation

a priori" (Kant 1781: A30). However, if time is the *a priori* form of inner intuition, then no comparison between experienced events and non-experienced events is possible since time only applies to the former. If we attempt to abstract objects from our sensory intuition then the objects lose the forms of intuition, which means losing their spatiotemporal nature (ibid.: B305). As such, it does not appear to make sense to attempt to draw up temporal relations between things existing outside and inside experience if we adopt a Kantian perspective. Temporal comparison would need to take place within experience, but as already discussed there is no vantage point available to either the subject or the experimenter to allow them to compare the time at which the subject's experience corresponds with anything else.

[4.5] Structural resemblance between events inside and outside consciousness

Before discussing any resemblance between conscious events and events lying outside of consciousness, it must be acknowledged that there are objections that such a thing is even possible. Berkeley objected on the grounds that a colour cannot be like something that is invisible, and something hard or soft cannot be like something intangible (Berkeley 1710: I.VIII). Locke and Russell again are excellent examples of philosophers who defended the view that there can be resemblance of a certain structural kind between the two kinds of events. In Locke's distinction between primary and secondary qualities, he stated that we must suppose that primary qualities, which are "extension, figure, number, and motion of bodies" must be actual properties of objects in order for them to be capable of affecting our senses (Locke 1689: II.VIII.12), whereas to explain how we perceive secondary qualities, such as colours, sounds and tastes (ibid.: II.VIII.10), we must only posit that an object can induce such sensations in us by virtue of their primary qualities affecting our senses (ibid.: II.VIII.13–14).

Russell argued that the evidence of our senses only allows us to attribute a certain mathematical structure to the external world:

> Colours and sounds can be arranged in an order with respect to several characteristics; we have a right to assume that their stimuli can be arranged in an order with respect to corresponding characteristics, but this, by itself, determines only certain logical properties of the stimuli. This applies to all varieties of percepts, and accounts for the fact that our knowledge of physics is mathematical: it is mathematical because no non-mathematical properties of the physical world can be inferred from perception. (Russell 1921: 253)

This led Russell to conclude that we can identify physical time with psychological time by virtue of them both sharing a structure "expressed by mathematical logic" (ibid.: 254). Both Locke and Russell determined that the world lying outside of our senses could not be compared to the world inside our senses in terms of how things appear or feel to us, but rather entirely in terms of structural properties that must be shared between our conscious states and the external world if the latter can cause the former. Indeed, Russell stated that the qualities of our percepts aside from general structural properties may not be determined by physical causes:

> That is to say, given the physical causal laws, and given enough knowledge of an initial group of events to determine the purely physical properties of their effects, it might nevertheless be the case that these effects could be qualitatively of different sorts. If that were so, physical determinism would not entail psychological determinism, since, given two percepts of identical structure but diverse quality, we could not tell which would result from a stimulus known only as to its physical, i.e. structural, properties. This is an unavoidable consequence of the abstractness of physics. (Russell 1927: 390)

Even if the qualities may be different between structurally identical conscious states, this still gives us an argument for the simultaneity of events, since we have two descriptions of events that share a certain structure, and this sharing of structure, we might argue, is best explained if our conscious states are caused by events in the external world.

There are many problems with this argument. For a start, our means of attributing this common structure to both worlds in the first place supposes that we can unproblematically refer to events inside and outside consciousness, which seems to run into some of the problems discussed above, such as the problems related to supposing that we can derive our sense of external time from that of subjective time. However, even if we could unproblematically arrive at the conclusion that there are structural similarities between conscious states and external events, this still would not suffice to demonstrate that they share a similar temporal structure.

To demonstrate this, it would need to be the case that we could posit a set of events in the external world causing our conscious states only if we state that they both run alongside one another in time. Yet, this is clearly not the case; we would in fact have a much greater degree of choice in the external events we must posit as causing our conscious states if we deny that they must be temporally simultaneous. If I assume that conscious states must be temporally related to their external causes in the way that Locke and Russell suggest, I could posit that the reason I experienced a bright flash in the sky followed by a loud noise is that there was electrical activity in the atmosphere that emitted light, picked up by my eyes a few moments before the sound waves produced by the resulting changes in air pressure were detected by my ears a few moments later. However, if I denied that the timing of my conscious state needed to match these specific external events at all then all I would need to posit is that any physical event causes the conscious state of perceiving a flash of light, such as a gust of wind blowing across my hair, and any other physical event causes the conscious state of hearing a loud sound, such as a piece of potato going into my mouth. These events could happen one after another, hours apart, weeks apart, or simultaneously. They could even

happen in either order without this necessarily affecting my experience of them. Indeed, if I reject the view that both occur alongside one another in time, there is such a huge variety of possible structures of external events that could correspond to my conscious events that it seems impossible to determine which events actually do so correspond.

As such, the only way to maintain that we can determine at least roughly which conscious events correspond to which external events is if we assert that it is impossible for conscious events not to share a temporal structure that is similar to that of external events. The problem is that there is no empirical means of refuting this claim given that a failure of the two sharing a temporal structure could *ex hypothesi* produce the same succession of conscious events. We would thus have to reject the possibility *a priori*, which would mean that we would already know *a priori* that conscious events share a temporal structure with external events. As in previously mentioned cases of accepting such *a priori* constraints, doing so here entails rejecting the possibility of a reliable empirical indicator for the timing of conscious events. We would instead be assuming that whatever our empirical evidence tells us, we could only infer from this evidence that certain external events are causing certain structurally similar conscious events, and as such we would be supposing that a science of consciousness cannot determine when conscious events occur.

[4.6] There can be no science of consciousness.

The above considerations have led to the conclusion that current consciousness science's task – to find empirical correlates of PC states – is doomed to failure and indeed all attempts have failed in their goals to this day. To recap, PC states are defined as conscious states that an individual can indicate the presence of through behaviour or reports. I argued in [2.5] that the criterion by which we determine if a state is a PC state – the capability to produce recognisable conscious behaviour – is so hopelessly vague that it risks making the

idea of PC states meaningless. I then argued in chapter 3 that, even if this criterion were somehow not so vague, the science of consciousness is firmly dependent upon the nonexistence of certain speculative entities that it does not have the means to disprove. No *a priori* objections to the existence of these entities can save the science of consciousness without doing equal damage to the prospects of an empirical study of consciousness.

Now we have ascertained that a science of consciousness has no prospect for even determining when a conscious state occurs. This reveals the impossibility of establishing what the empirically observable correlates of even PC states are.

[4.7] Summary

Assessing Libet et al.'s cases of subjective delay, we found several reasons for philosophical disagreement with the researchers' conclusions [4.1]. These conclusions rely on a flawed demonstration of the timing of conscious states, assuming without evidence or argument that we can ascertain the timing of events lying inside and outside consciousness and compare them. We presented possible justifications for such an assumption but ultimately showed that none of the positions presented offered us reason to think that such comparisons are possible [4.2] – [4.5]. This led us to the conclusion that a science of consciousness is impossible [4.6].

In the next chapter, we will be introduced to the Consciousness Science Paradox, and will begin to see the impact this paradox has on our understanding of the nature of scientific enquiry.

5

Introducing the Consciousness Science Paradox

[5.1] Two contradictory claims

The Consciousness Science Paradox consists of two seemingly contradictory claims:

Claim A: PC states must, in principle, be possible to identify through standard scientific means such that statements about consciousness must be possible to verify or falsify on the basis of observation.

Claim B: PC states cannot, in principle, be identified through standard scientific means and so statements about consciousness are not possible to verify or falsify on the basis of observation.

The first claim was established by ascertaining that conscious states must play a causal role in our production of behaviour indicating their presence. The second claim was established by analysing scientific discourse about consciousness and showing how it necessarily fails to give us any means of empirically determining where or when the conscious states in question are present and where or when they are absent. This is, I believe, a significant philosophical problem for many metaphysical positions for which both claims are demonstrably held to be true.

[5.2] What are "consciousness scientists" actually studying?

Before beginning to evaluate which metaphysical perspectives are affected by this paradox, the question must be answered: if "consciousness science" is not studying consciousness, then what is it studying?

I think it would not be appropriate to claim that it is studying nothing, even if there are examples where conclusions are arrived at beyond what the evidence suggests. Tallis has pointed to several such examples where it has been declared by researchers that the neurophysiological correlates of various things such as love and beauty have been found on the basis of flawed experimental methods, such as by looking at fMRI scans when an individual is looking at a picture of somebody they love, or at a painting the individual has said is beautiful -- as if the ideas of love or beauty are of things that wholly exist in a single instance of looking at a picture (Tallis 2011: 74–75). He also pointed out that fMRI scans detect significant changes in blood flow, which only occur where activity is concentrated in a particular area of the brain, and so activity more dispersed across the brain would be less likely to show up on such scans (ibid.: 76).

It is also true, as we have explored, that the claims of consciousness science cannot relate to the subject matter in the way prominent advocates of such an approach believe it to. The claims these scientists are making about consciousness, in fact, are likely false given that their methodology does not give them any means of supporting those claims.

Nonetheless, the study seems for the most part to be making a form of progress. Claims appear to be falsified by the results of studies, and there appear to be common clear research directions that are adopted by many scientists. This can all be accounted for without mention of consciousness.

To study the empirically observable correlates of consciousness, scientists have been correlating the presence of recognisable conscious behaviour with the empirically observable states that produce those behaviours, such as physiological and cognitive states. As such, it

would be more accurate to claim that scientists are studying the physiological and cognitive states responsible for recognisable conscious behaviour, although as we shall see even this is a problematic interpretation.

The empirically useful content of consciousness science could easily be partitioned from the claims that essentially mention consciousness. By "empirically useful," I mean to distinguish the claims that are supported by empirical observations, as some of these scientists also make claims that seem rather more speculative and could be removed entirely from the discourse without this seriously altering the direction of scientific research. I present a couple of examples below.

> It is difficult for many people to accept that what they see is a symbolic interpretation of the world – it all seems so like "the real thing." But in fact we have no direct knowledge of objects in the world. This is an illusion produced by the very efficiency of the system since, as we have seen, our interpretations can occasionally be wrong. Instead, people often prefer to believe that there is a disembodied soul that, in some utterly mysterious way, does the actual seeing, helped by the elaborate apparatus of the brain. Such people are called "dualists" - they believe that matter is one thing and mind is something completely different. (Crick 1994: 33)

This unusual paragraph in Crick's "The Astonishing Hypothesis," seems to indicate that he was not a direct realist and that he believed the alternative to direct realism to be substance dualism. This is quite puzzling for anybody with knowledge of these philosophical positions, since whether one is a direct or indirect realist does not seem to self-evidently entail either the truth or falsehood of substance dualism. Certainly, we should wish to hear some argument that it does. In any case, this claim seems to be irrelevant to the sort of methodology he favoured for studying consciousness; he could have arrived at his association between

consciousness and reverberatory circuits running from the thalamus to certain cortical layers without such speculation.

Ramachandran also made claims that seemed to be abstracted away from the content of empirical research:

> Obviously self and qualia are two sides of the same coin. You can't have free-floating sensations or qualia with no one to experience them and you can't have a self completely devoid of sensory experiences, memories or emotions… What exactly is meant by the "self"..? First of all, continuity: a sense of an unbroken thread running through the whole fabric of our experience with the accompanying feeling of past, present and future… Fourth, a sense of agency, what we call free will, being in charge of our own actions and destinies… Fifth, and most elusive of all, the self, almost by its very nature, is capable of reflection - of being aware of itself. A self that's unaware of itself is an oxymoron. (Ramachandran 2004: 96-97)

The points where any of these claims have touched on empirical content are very hard to distinguish. The notion of the self is largely defined according to some reflections Ramachandran has had, such as that a self cannot be unaware of itself, but it is not clear that this touches on any specific empirical model of the "self." Indeed, the claim that a self being unaware of itself is an oxymoron suggests that Ramachandran took it to be *a priori* or analytically true that a self necessarily is aware of itself. Equally, the relationship between the self and qualia does not appear to be something that a science of recognisable conscious behaviour should hope to be able to determine, nor that it should be required to have a stance on.

My recommendation is that speculations such as those just mentioned by Crick and Ramachandran are discarded entirely from the scientific enterprise they have commented on and contributed to. Science may rely on theoretical entities and processes to explain evidence, and they may use concepts adapted from philosophical notions of consciousness, but only insofar as these concepts are useful in the explanation of evidence. Speculation seen in the above examples is not necessary for such explanations and does not constitute part of any empirical content of current scientific investigation that can be partitioned from the current problematic attempts at a science of consciousness. Of course epistemological and metaphysical views may form the basis for a rational justification of any existing scientific approach, but many prominent scientists in this area including Crick and Ramachandran cannot be said to have been interested in such projects given their bold interventions into philosophical issues with no effort to engage with the relevant literature. Additionally, their works are often peppered with disparaging and condescending remarks about philosophical efforts to understand the nature of consciousness. I have given several such examples below:

> The next thing to stress is that the study of consciousness is a scientific problem. Science is not separated from it by some insurmountable barrier. If there is any lesson to be learned from this book it is that we can now see ways of approaching the problem experimentally. There is no justification for the view that only philosophers can deal with it. Philosophers have had such a poor record over the last two thousand years that they would do better to show a certain modesty rather than the lofty superiority that they usually display. (Crick 1994: 257-258)

> I will firmly assert, however, that these philosophical arguments, which rely so heavily on abstract logic for ammunition, as they neglect the scientific

enterprise, provide very limited insights into consciousness, and can be positively misleading. (Bor 2012: 4)

It is my belief that such a deeper understanding of the brain will have a profound impact not just on the sciences but on the humanities as well. Lofty questions about the mind are fascinating to ask – philosophers have been asking them for three millenia both in my native India and in the West – but it is only in the brain that we can eventually hope to find the answers. (Ramachandran 2004:39)

There is indeed a fundamental limitation on philosophical efforts to discern the origins of consciousness that arises, in part, from the presumption that the sources of conscious thought can be revealed by thinking alone. The presumption is as patently inadequate as efforts in previous times to understand cosmogony, the basis of life, and the fine structure of matter in the absence of scientific observations and experiments. (Edelman & Tononi 2000: 6)

Don't be taken in by philosophical grandstanding and proclamations that the Hard Problem of consciousness will always remain with us. Philosophers deal in belief systems, simple logic, and opinions, not in natural laws and facts. They ask interesting questions and pose charming and challenging dilemmas, but they have a mediocre historical record of prognostication... Listen instead to Francis Crick, a scholar with a far better track record of predictions: “It is very rash to say that things are beyond the scope of science.” (Koch 2012: 137-138)

Given the absence of engagement with philosophy by such scientists, sweeping statements about the nature and limits of philosophy do not seem to be within their remit of authority. At best, they could be said to be authoritative over the questions within their particular research areas, but insofar as these questions touch on questions about the limits of science and the boundary between science and philosophy, these are also questions that require actual engagement with philosophy of science rather than a dismissive wave of the hand.

Although I have attempted to separate the study of consciousness from the study of recognisable conscious behaviour here, the distinction may seem inadequate on the grounds that the idea of recognisable conscious behaviour seems dependent upon the concept of consciousness; a behaviour is only recognisably conscious if consciousness somehow underpins it. This seems to me a reasonable assessment.

The scientific study of consciousness instead either corresponds with the scientific study of some empirically observable cognitive or physiological state, such as sufficient information integration or global availability, or it fails to match up with any empirically observable state and is rather a group of studies with some crossover of content that have simply fallen under the same ideological heading, such as the study of the capacity to discriminate stimuli non-verbally, the study of the effects of general anaesthesic on the central nervous system, and the study of the surgical division of hemispheres. It will be up to the progress of such areas in science to determine specifically which concepts are most appropriate and there may not be a single concept that will suffice to explain all that currently falls within the remit of consciousness science, but if there is it will be a concept that is required in order to explain the evidence, rather than a concept that requires more than can be explained by the evidence.

Some examples of such concepts are attention, global availability, physiological responsiveness and awareness. Global availability is measured by the capacity of various systems to utilise the same information, physiological responsiveness can be measured by the

presence of certain physiological activity given certain stimuli and attention can be measured by the capacity of an empirically observable state to discriminate certain stimuli.

The reason I include “awareness” in this section, even though it may appear to straddle a line between scientifically useful concepts and the concept of consciousness I am seeking to partition the content of this research from, is that there is a scientifically useful concept of awareness that can be distinguished from the use of the word meaning specifically *conscious* awareness.

Block pointed out that we can be “phenomenally conscious” without being “access conscious” (Block 1995: 233–236). By this he means that there are cases where we appear to be aware of something but are unable to use our awareness of that thing to produce a response or behaviour. For example, we may be involved in an intense conversation and then notice a drilling noise coming from outside that we recall having been aware of for some time. Block stated that in this case we appear to be aware of the noise while, in an important sense, not being conscious of it (ibid.: 234). Similarly, neuroscientists Tononi and Koch determined that you could have consciousness in the absence of attention because there are cases where a subject’s attention is occupied by some task but they can still make accurate judgements about stimuli being presented alongside that task (Tononi & Koch 2008: 242). The trouble in both of these cases is that these phenomena could acceptably fit into a scientific study that did not claim to be about consciousness; instead what they are describing is some particular means of processing information that produces observable effects, such as your capacity to accurately report that such-and-such stimuli were present, but that does not fit a concept such as attention. Personally, to avoid sliding into ambiguous talk about consciousness and running into the myriad problems mentioned above, I think it would be better to avoid naming such a mechanism after any sort of consciousness or awareness. Nonetheless, whatever we should name such mechanisms, they can be described in empirically unambiguous terms; awareness in this sense is nothing other than the capacity to process certain information in the absence of attention.

Given that there is no requirement to consider the study of such phenomena a study of consciousness, *I shall hereafter cease to refer to any current scientific study as "the science of consciousness."*

[5.3] The Paradox as a philosophical problem

The Consciousness Science Paradox is a philosophical problem arising from competing descriptions of what should be possible to achieve through empirical investigation with regard to the study of PC states. If I have successfully defended these competing descriptions, then it is a serious problem given that PC states are the only possible kind of conscious states that could be empirically studied in principle; if there are disembodied conscious states or conscious states that produce no observable effects then of course these would fail to admit empirical investigation. Showing that even PC states are philosophically problematic tells us that our reasoning must be mistaken somewhere.

My goal is to show that this is a philosophical problem worthy of attention. While problems about various matters related to consciousness are prevalent in philosophical discussion, such as the existence or non-existence of qualitative properties and the relation of consciousness to the physical world, these discussions generally take it for granted that the possibility of identifying empirical correlates of consciousness is not itself problematic.

There are two immediate responses that one could give to defuse my arguments, and these are:

1) This problem only applies given a certain conception of science.

2) This problem only applies given a certain conception of consciousness.

While I cannot show that no conception of science or consciousness would avoid my paradoxical conclusion, I can show that the paradox is not narrowly constrained to a small subset of philosophical positions in both of these areas. In fact, even major revisions to our conceptions of consciousness and science do not give us a way of avoiding the paradox. I will first demonstrate the problem with attempting to modify our conception of science.

[5.4] Evaluating our conception of science

There are many different philosophical viewpoints on the nature of science that would all give different answers to the question, "What are the essential features of a scientific claim as opposed to a non-scientific one?" Popperian falsificationism demarcates science purely by the falsifiability of the claim; if a claim is falsifiable through observation it is scientific, and if it is not falsifiable through observation then it is not scientific. About as far from this viewpoint as possible is the anarchism of Feyerabend, which claims that there is nothing to demarcate scientific claims from non-scientific claims since there is no scientific method accepted from all vantage points across all time.

If we believe that it is only particular viewpoints on the nature of science that are vulnerable to the paradox, we may seek to adopt a perspective for which the contradiction would not arise. We can try to avoid the paradox using a new conception of science by doing one of two things:

a) **Attempting to conceive of science as being able to explain consciousness**
b) **Attempting to conceive of science as not being required to explain consciousness**

I will evaluate both of these possibilities in turn. It is no trivial task to determine which perspective on the nature of science is correct and I have no strong opinion on the subject. As such I will simply survey perspectives here to demonstrate that the conclusion I wish to draw is not restricted to a particular viewpoint in the philosophy of science.

[5.5] Attempting to conceive of science as being able to explain consciousness

The idea that there is a common conception of science is disputable. There are multiple competing views on the nature of science that would have something different to say about what sort of modifications are possible. I will take a representative sample of such positions in order to demonstrate that no matter which we take, each are equally incapable of studying consciousness.

[5.5.1] Falsificationism

The goal of falsificationism was to present a view of science by virtue of which it could both be demarcated from non-science or pseudoscience and be described using deductive logic rather than inductive logic (Popper 1935: 9–12). The essential characteristic of a scientific system under such a view is that "*it must be possible for a scientific system to be refuted by experience*" (ibid.: 18).

By "experience," Popper referred not to individual subjective experience but to things that can be "*inter-subjectively tested*" (ibid.: 22). And for a system to be falsifiable is for it to rule out "*at least one event*" (ibid.: 70). Now, whether the correlative approach to the study of consciousness is falsifiable in this regard is disputable. If, in blatant disregard of the criticisms of the previous chapter, I assume that the presence of the conscious state of feeling pain at

time X is established by an individual stating, “I feel pain,” at time X, and that the absence of that conscious state at time Y is established by the individual stating, “I do not feel pain,” at time Y, then I will agree that the claim, “The individual is in the conscious state of feeling pain at time X,” is inter-subjectively testable and rules out certain events, such as that the individual is not in that conscious state at X and is in that conscious state at Y.

The view that Popper’s falsificationism is simply the view that theories are only rejected if they are falsified has been described as “naive falsificationism.” Kuhn, who I will discuss shortly, was accused by Lakatos of conflating naive falsificationism with Popper’s actual view in his criticisms of Popper (Lakatos 1970: 10). To see the problem with naive falsificationism and why it is inconsistent with Popper’s project, I will assess the attempt to study consciousness scientifically using it.

Popper’s description of falsificationism is primarily a story about how scientific progress is made and why one theory is replaced by another and in this sense I should be asking whether such a story could tell us why such scientists would be inclined to drop mention of consciousness. This is because there is no explicitly mentioned theoretical move to associate, for instance, the capacity to produce certain behaviours with the presence of consciousness, but rather it it implicitly made where this capacity is an object of scientific study. So, statements regarding the presence of consciousness, even if they conflate consciousness with the capacity to produce certain behaviour, are falsifiable and have added new empirical content to the studies of cognitive science and neuroscience.

Our goal here is to know whether it is worth dropping one explanation in favour or another according to falsificationism as characterised so far. Consider:

1) Information integration of a certain minimum degree (D) is required for human consciousness.

2) Information integration of a certain minimum degree (D) is not required for human consciousness but is required for the capacity to produce behaviour associated with consciousness.

The second statement adds no new falsifiable potential that is not included in the first because any claim about consciousness is only verifiable or falsifiable by the extent to which the capacity to produce behaviour associated with consciousness is affected. There is no empirical way to distinguish between claims of the first sort and the second sort, and so the distinction between the two cannot be made in terms of falsifiable content. No new observations have become available for falsification as a result of moving from the first statement to the second. As such, scientists should equally prefer both positions.

Our concern should be that whether we accept that consciousness is present or absent, the falsifiable content of scientific theories remains the same. For the removal of consciousness from scientific discourse to result in no removal of empirical content tells us that the only function it serves is that of an historical accident; if we had instead conflated the capacity to produce recognisable conscious behaviour with the presence of angels we would have the same empirical content. As such, the defence we are attempting to muster here is that scientists have no reason to abandon their descriptions of consciousness because whether they mention consciousness or do not is irrelevant to their scientific work. This would be a weak reading of falsificationism because it would entail that there is no way to demarcate a theory that is falsifiable from one that contains the same falsifiable content but also has many unnecessary supplementary hypotheses that add no falsifiable content. This clearly goes against Popper's task to demarcate science from non-science.

If I claim that you will have good fortune involving the letter "p" over the next month as a result of a specific movement of certain celestial bodies, then this claim is falsifiable because certain events are ruled out, such as that you will have only bad fortune involving the letter "p," and is inter-subjectively testable in that multiple individuals could verify the same movement

of celestial bodies. The reason that the pseudoscientific statement about celestial bodies can be demarcated from a genuine scientific statement is that it contains auxiliary hypotheses that have no effect on the overall falsifiability of the theory, which Popper found unacceptable (Popper 1935: 62). If I claim that a certain celestial body will align with a certain other celestial body at a certain time, then this is something that would be falsified if we were to observe the two bodies at that time not in alignment. If I claim that the two celestial bodies will align at that time and that as a consequence you will have good luck involving the letter "p" over the next month, this adds no additional observations we could make to falsify the overall theory; there are no clear observations that could confirm that somebody has had good or bad luck in a way that relates to a particular letter. Yet, since it has a greater number of auxiliary hypotheses it does stand a greater chance of clashing with reality (ibid.: 272). Because of this, the former claim is more scientifically acceptable than the latter.

The same applies with contemporary efforts to identify empirical correlates of consciousness. The claims that are made of the type, "Neurophysiological state X was observed simultaneously with the subject exhibiting behaviour Y," are more scientifically useful than claims such as, "Neurophysiological state X was observed simultaneously with the subject being in conscious state Z," since the latter claim requires claims of the former sort to be true in addition to auxiliary hypotheses such as that conscious states correspond with certain behavioural states, which are not falsifiable. We should, then, prefer to make the former sort of claim because it is less likely to clash with reality and if we agree with Popper we should adhere to his principle of parsimony (ibid.: 131).

[5.5.2] Scientific revolutions

For Kuhn, scientific progress depended on the ability for scientists to rely upon existing paradigms that they do not need to provide an explanation of:

> When the individual scientist can take a paradigm for granted, he need no longer, in his major works, attempt to build his field anew, starting from first principles and justifying the use of each concept introduced. That can be left to the writer of textbooks. (Kuhn 1962: 20)

"Normal science" for Kuhn was the attempt to describe nature purely in the terms allowed by the paradigm (ibid.: 24) and consisted wholly of three classes of problem:

1) Determination of significant fact
2) Matching of facts with theory
3) Articulation of theory (ibid.: 34)

Where normal science ended was where phenomena appeared that could not be understood within an existing paradigm. Kuhn referred to such phenomena as "anomalies" (ibid.: 52). He argued that anomalies could be found and could contribute to putting a theory in a state of crisis (ibid.: 66-76). Kuhn then stated that the mere presence of an anomaly was insufficient to put a theory in crisis because the response to these may be simply to make further *ad hoc* assumptions that would explain away such anomalous results (ibid.: 78). Nonetheless, depending on the paradigm and the historical context, the presence of anomalies could put a theory into a state of crisis, at which point ordinary science would transition to extraordinary science (ibid.: 82). Within extraordinary science, explaining the anomaly becomes the main research project. This could be done either by the scientific community fitting the anomaly into the original paradigm, setting the anomaly aside for future research, or by the scientific community endorsing a new paradigm within which the anomaly can be explained (ibid.: 84). Kuhn wrote:

> It is, I think, particularly in cases of acknowledged crisis that scientists have turned to philosophical analysis as a device for unlocking the riddles of their field… It is no accident that the emergence of Newtonian physics in the seventeenth century and of relativity and quantum mechanics in the twentieth should have been both preceded and accompanied by fundamental philosophical analyses of the contemporary research tradition (ibid.: 88).

The criticisms of the above efforts to produce a science of consciousness could be seen, if we were to adopt a Kuhnian standpoint, as a project to demonstrate that conscious states are anomalies. This could then push science toward a period of extraordinary scientific research whereby new ways to explain the anomalous results should be sought.

This project is a philosophical analysis of the contemporary research tradition and as such could fit the role Kuhn described. The trouble of course is that we are not in a period of acknowledged crisis; the literature written by scientists on their efforts to understand consciousness, while often acknowledging difficult philosophical questions, seems to express an optimism among scientists regarding the current research projects and a lack of acknowledgement of fundamental difficulties in their approaches. The case I have presented is where I point out that the central phenomena that the current approach seeks to explain is necessarily anomalous; it will always lie outside the scope of this project to explain how or why a conscious state should be related to a particular empirically observable state because the relationship between that state and the behaviour that we use to detect it is scientifically inexplicable.

For instance, the current paradigm must accept that consciousness has an effect on our capacity to produce certain behaviour, but this acceptance itself assumes that consciousness must be observable by somebody at least, even if just the individual experiencing their own conscious states, and that consciousness must have observable effects despite the fact that the tools this paradigm uses to understand the phenomena it seeks to

explain, such as psychology experiments and observations of the activity of neurological structures, shed no light on what these observable effects are. Indeed, they cannot shed light on what these observable effects are, since they must be assumed for such an investigation to begin in the first place.

Furthermore, efforts to produce a correlative science of consciousness are hailed as efforts to understand consciousness using the tools of neuroscience and cognitive science, where all of the established research in the field involves the use of these tools, and so it is a crucial feature of the current paradigm that such an approach is a subset of the overall disciplines of neuroscience and cognitive science. As such, it is problematic for such a study to require the additional *ad hoc* assumptions regarding an essential and indisputable spatiotemporal correspondence between conscious states and recognisable conscious behaviour, which is required for the correlative approach to work, when no such assumptions are required for the rest of neuroscience and cognitive science. This makes it clear that the science of consciousness is not a proper subset of these fields, which makes the phenomena it seeks to study anomalous within those overall scientific projects.

Perhaps it will be denied that the phenomena of consciousness are anomalous because they are not observed at all and are rather derived from certain other observable facts. But this denial takes us no closer to a resolution of the Consciousness Science Paradox. Even if derived from other observable facts, for consciousness to exist it must have observable effects, as illustrated in the earlier refutation of epiphenomenalism. As such, the effects it produces must be observable, and if we can thus derive the existence of consciousness from these effects, it still has the status of an entity that we need to posit to describe our empirical observations. Even in criticising consciousness science, we are not required to deny that conscious states only produce observable effects; we have only found problematic the assumption that conscious states must be able to produce effects that we would recognise as belonging to conscious states, such as recognisable conscious behaviour. There is thus no

reason that consciousness should be inappropriately labelled an anomaly if it fails to adhere to scientific explanation within a given paradigm.

A pessimistic interpretation of my analysis of current efforts to study consciousness science would be to state that, even if scientists were to recognise conscious states as anomalous, they would simply relegate the anomaly to future science to resolve. Yet, this would clearly be woefully inadequate as the phenomena seen as anomalous here are the phenomena most central to this scientific project. It would be rather like claiming to study an astronomical entity while shrugging off the fact that the effects of that entity are *in principle* impossible to identify. Nonetheless, the Kuhnian picture does not state that scientists necessarily acknowledge where their paradigms break down, it merely attempts to describe what happens when they do.

[5.5.3] The Bayesian approach

The Bayesian approach relies upon the use of theorems developed from the work of Thomas Bayes, in particular a posthumously published letter where he posited a solution to the problem of determining the probability of an event being observed that has not been observed over multiple previous trials (Bayes 1763). Howson and Urbach gave three variations of Bayesian theorems with the second and third variants being developed after Bayes' initial formulation (Howson & Urbach 2006: 20–21), but I will only give the first here, the one developed by Bayes, as a demonstration of what Bayesian theorems are designed to describe.

A Bayesian Theorem:

$$\boldsymbol{P(h|e)=\frac{P(e|h)P(h)}{P(e)}}\text{, where } P(h) \text{ and } P(e) \text{ are greater than } 0.$$

The letters *h* and *e* stand for hypothesis and evidence, and the letter *P* stands for the probability of something occurring. Let us assume that we wish to ascertain the likelihood that somebody has a specific disease given a positive test result. On the left side of the theorem is the probability that our hypothesis (that this subject has the disease) is correct given the presence of our evidence (the positive test result) [*P(h | e)*]. This is what we wish to ascertain, which we do by dividing the two figures on the right side of the theorem.

The first figure we get by multiplying two elements, which are found on the top half of the right side of the theorem. The first element is the probability that we would receive a positive test result if the hypothesis is correct [*P(e | h)*] (this is the likelihood of somebody with the disease receiving a true positive result). Let us say this probability is 75%; if you have the disease there is a 75% chance you will receive a true positive. The second element is the probability that the hypothesis would be correct regardless of the test result [*P(h)*] (this is the overall likelihood of any random person in the population having the disease). Let us say that the probability of anybody having the disease in our chosen population is 15%. Multiplying these two probabilities together gives us the probability of a true positive result in any random member of the population (which in this case would be 11.25%).

The final figure we need to find is the odds of the probability of a positive result in any random member of the population [*P(e)*], thus including false positives. Let us say a positive result is 20% likely.

This means that, in order to determine what the probability is of an individual having the disease given a positive result, we must divide the probability of a true positive result by the probability of any positive result, which gives us 56.25% in this example. The probability, then, that our individual with a positive test result has the disease is 56.25%.

Expected features of science appear to be borne out by this theorem. For instance, the presence of a given piece of evidence over a single trial greatly increases the probability of the hypothesis being correct, but subsequent instances of that piece of evidence being present increase that probability to a smaller and smaller degree. So, a theory completely unsupported by evidence at a given time will then be seen as much more likely to be correct if that evidence is observed, but a theory that has many results confirming it will not be significantly impacted by the presence of a further single observed instance. This "law of diminishing returns" (Popper 1962: 240) is expected because our scientific understanding appears to be less affected by an observation of something we have already observed on multiple occasions than by a novel observation, with the latter having a more significant impact on our theoretical understanding of reality.

Bringing this back to discussion of consciousness, it can easily be made clear that Bayesian theorems give us no tools by which we can establish that theories about consciousness in particular are amenable to scientific investigation. For instance, let us envision what it would look like to evaluate the possibility that a seemingly comatose individual is actually conscious at a given moment. Let's say that 0.5% of individuals in a seemingly comatose state are actually people with locked-in syndrome. Let's say that locked-in syndrome, where present, is correctly identified 80% of the time. Finally, let's say that the odds of a seemingly comatose individual being identified as having locked-in syndrome (correctly or incorrectly) is 1%. By these numbers, the odds of somebody who is identified as having locked-in syndrome *actually* having it are 40%.

If these figures were correct (although they almost definitely are not), what do they tell us about consciousness? Well, first we would need to establish that consciousness is absent in comatose patients not suffering from locked-in syndrome and present in patients with locked-in syndrome. As such, we would first need to ascertain criteria by virtue of which consciousness can be identified and, as already described in detail, this can only be done by assuming that the presence of physiological states underlying certain behavioural capacities

is associated with consciousness. This means that, given that the extent to which it is possible to identify that an individual has locked-in syndrome is identical to the extent to which it is possible to identify that a certain physiological state associated with a particular behavioural capacity (in healthy participants at least) is present, the odds of us ascertaining that a seemingly comatose individual who has been identified as having locked-in syndrome is actually conscious will also be 40%. It is clear, then, that claims about consciousness are doing no work in this example; the numbers being calculated are done so purely on the basis of the presence of states underlying behavioural capacities by virtue of which the presence of consciousness is ascertained. The numbers themselves tell us nothing about whether or not it is correct to ascertain the presence of consciousness in this way.

This can be further demonstrated by enquiring into how we could assess the possibility that consciousness is present where the capacity to produce recognisable conscious behaviours is present. In this case, [$P(h \mid e)$] would be the odds that our hypothesis, that consciousness is present in an individual, is true given the evidence, that recognisable conscious behaviour is present in that individual. It should be clear that, in attempting to find this value, there are certain initial values that could not be ascertained. How, for instance, would we ascertain the value of the probability that the hypothesis is true across the entire population [$P(h)$]? We could only do so by the prevalence of recognisable conscious behaviour, and as such we would have to beg the question; one of the values being used to calculate [$P(h \mid e)$] could only be known *after* this value has been calculated, which is clearly circular.

We might be able to use the Bayesian theorem to calculate how often we correctly identify that recognisable conscious behaviour is present in individuals that exhibit it, but in this case the presence or absence of consciousness would be doing no work toward the calculation of any of the figures used in Bayesian theorems; it would simply be the presence of the behavioural capacity that would be used for the calculation.

There are multiple variations on Bayesian theorems, but they generally use elements that are not significantly distinct to those above and so the identity between the figure for identifying the presence of locked-in syndrome and identifying the presence of consciousness is seemingly comatose patients would be replicated, which means that statements about the presence or absence of consciousness are doing no work in producing any of the figures. The three forms outlined by Howson and Urbach (Howson & Urbach 2006: 20–21) all use the element [*P(h)*] and so the problem outlined above in ascertaining the probability that consciousness is present in an individual exhibiting recognisable conscious behaviour would be replicated in all such variations on the theorem. This shows that statements about consciousness are either irrelevant to Bayesian theorems or, worse, are unworkable within them.

[5.5.4] Summary of attempts

I have given only a small sample of perspectives that would each say something different about the nature of science, none of which have given a means of conceiving of science in such a way that it becomes able to study consciousness. Briefly, falsificationism left us with the problem that we should prefer a theory that simply posits a relationship between empirically observable states without mentioning consciousness because it is more parsimonious, Kuhn's revolutionary view opened the door to the possibility of entering a period of extraordinary science whereby explanations are sought for the anomalous results of current attempts to study consciousness, and Bayesian theorems were shown to be incapable of factoring in claims about the presence of consciousness in a meaningful way.

Where there is overlap between all these positions is in the agreement that it is harmful for a science to rely upon hypotheses that cannot be empirically demonstrated to be either true or false. Either this would make our scientific hypotheses less falsifiable, would present them

with anomalies, or would make the truth of our hypotheses given the evidence unlikely. There is a clear reason why this is not just related to the positions under discussion, but that it must be the case.

If we are to assume that there is a way of theorising about science such that we would posit that consciousness could be made explicable within scientific terms, we must suppose that there is a link between evidence and hypotheses such that, given certain evidence, certain hypotheses about consciousness will turn out to either be true or false. Yet, as the discussion of the correlative approach to a science of consciousness above has shown, there is no evidence that could support or contradict basic claims about the relationship between consciousness and certain physiological, cognitive or behavioural states.

If we are to assume that there is some way we can describe a link between evidence and hypotheses such that it will become explicable for us to demonstrate that, for instance, conscious events can be observed to be correlated with events occurring outside consciousness, we would be supposing that there is some way we can separately observe what is inside and outside consciousness. This is, quite clearly, not a redescription of the process of science, but is rather a redescription of the nature of consciousness.

While that which constitutes a piece of evidence may differ between the aforementioned positions, it is quite clear that the evidence must be couched in observational terms. As mentioned earlier in discussion of Feyerabend, I do not pretend to know what the sufficient conditions are for a discipline such that it may be considered "scientific." Nonetheless, a science must be able to distinguish between facts on the basis of observations, even if what is considered a "fact" and an "observation" may also be unclearly defined. This means that a science cannot support a claim that is incapable of being demonstrated or refuted by observed facts. At best it can deny the importance of explaining such facts, as will be demonstrated in the alternative modification to the conception of science where we suppose that science should not be expected to explain consciousness. I do not claim either that a science must rely solely upon empirically verifiable or falsifiable claims, but rather that if we

are to claim that consciousness is scientifically explicable, we must be conceiving of science in such a way that it can empirically distinguish between true and false claims about consciousness, which has been demonstrated to be impossible. Indeed, if we regard it as given that consciousness is related to particular cognitive capacities or behavioural states, our assumption must be that a science of consciousness does not need to provide an explanation of how consciousness is known to be related to such empirically observable states. This is the second possibility with regard to modifying our conception of science, which I will now address.

[5.6] Attempting to conceive of science as not being required to explain consciousness

There are two approaches I will examine that could maintain that science has no requirement to be able to study consciousness, and thus we may seek to avoid the paradox by adopting one of these two approaches. We will examine the research programmes view of Lakatos and we will revisit the anarchism of Feyerabend.

[5.6.1] Research Programmes

Lakatos argued against Kuhn, stating that Kuhn defended his position by espousing its benefits over a weak form of naive falsificationism (Lakatos 1970: 10). Instead, Lakatos argued in favour of a form of sophisticated falsificationism where a theory can have a set of core commitments that may not be admissible to direct empirical falsification but that enable us to posit auxiliary hypotheses that can clash with other basic statements regarding singularly spatiotemporally locatable events (ibid.: 41). As such, science does not discard a research programme focused around one theoretical core every time its statements are contradicted by

observation. Rather, a main theoretical core persists until it can be replaced by another which has auxiliary hypotheses with greater empirical content (ibid: 41–42).

Lakatos even allowed that this core could be ‘metaphysical’ rather than ‘refutable’ without this resulting in a difference in the methodology adopted by that research programme (ibid.: 42).

Lakatos stated further:

> The problem is not what to do when 'theories' clash with 'facts'. Such a 'clash' is only suggested by the 'monotheoretical deductive model'. Whether a proposition is a 'fact' or a 'theory' in the context of a test-situation depends on our methodological decision. (ibid.: 44)

Lakatos stated that the actual clash in such a situation is between our interpretations of which theory provides the facts and which provides an explanation of the facts. The goal then is to attempt to replace each theory, or even both, to see if the new theoretical position leads to "the biggest increase in corroborated content" (ibid.: 45).

The “hard core” of a research programme is the set of theoretical commitments which the programme forbids us to refute, and the auxiliary hypotheses are its protective belt, which are subject to replacement and refutation (ibid.: 48). Lakatos gave an example of how a new research programme could come to replace an old one, with Copernican astronomy coming to replace Aristotelian physics first by being “grafted” onto the Aristotelian research programme despite the clear inconsistencies between the two positions and then, as Copernican astronomy gradually grew in explanatory power, the relationship between the two research programmes ceased to be symbiotic and became competitive instead (ibid.: 56–57).

His main objection to Kuhn's revolutionary view on science was that it assumed that scientific research programmes go through long periods of being the only accepted programme, only to eventually be replaced in a scientific revolution (ibid.: 69). Instead,

research programmes are those that can explain "the previous success of its rival" and that supersede their rivals "by a further display of heuristic power" (ibid.: 69). “Heuristic power” in this context refers to the capacity of the programme to predict novel facts.

Under Lakatos’ view on research programmes, it is arguable that there is no particular need for science to fully be able to justify core beliefs. As such, a research programme to understand consciousness could have core beliefs about what constitutes an observation of a conscious state, the timing of that conscious state and the relationship between a conscious state and external non-conscious states and could have no means of justifying these beliefs. Instead, these beliefs could be justified by non-scientific reasoning, or could perhaps have no further justification at all beyond their relationship to other claims that are amenable to scientific verification or falsification. The only requirement for a set of beliefs to form the hard core of a research programme is that they enable us to posit hypotheses that are amenable to scientific investigation.

There is one argument one could make against such a justification of correlative approaches to the scientific study of consciousness that I think is indecisive but that is useful to consider. It could be argued that if consciousness is unexplained under such a view, then whether it forms the core of a scientific research programme is irrelevant for our understanding of consciousness. After all, the purpose of a science of consciousness in the first place is to understand how consciousness fits into the natural world, and so a view on science that describes a certain scientific approach as simply taking its relationship to other natural phenomena for granted does not give us any means of fulfilling that purpose.

The trouble with this position is that it describes how research programmes would leave our claims about consciousness unjustified, whereas Lakatos was not attempting to justify the claims forming the core of a research programme but was rather attempting only to give a description of the process by which scientific theories are rejected and accepted. If Lakatos’ description was accurate, the question as to whether beliefs making up the hard core of a research programme are justified by the success of that research programme may not admit

an obvious answer, but we could make a case that they are. After all, if the role of a scientific enterprise is to give us an accurate description of certain aspects of reality, and a scientific enterprise will always contain core claims that enable us to posit testable hypotheses, then it seems plausible that the success of those descriptions will be partially dependent upon those core claims by which they could be made. If I have a description of reality that has been established by a certain observation but that is dependent upon the truth of another claim that cannot be directly supported by observation, it seems reasonable to state that both the empirically demonstrated claim and that claim which it is dependent upon are supported by that observation.

This justification does not appear to support the correlative approach to a science of consciousness though because, as I have demonstrated, all claims about consciousness have a component that is amenable to empirical investigation, such as that a physiological state underlies a particular capability to produce certain behaviour, and a component that is not, such as that consciousness is correlated with the capacity to produce certain behaviour. This means that the success of the auxiliary hypotheses that follow from our core claims about consciousness cannot support such core claims given that the same success of such hypotheses could occur without a core that mentions consciousness at all.

While this may be so, this still does not give us a direct means of rejecting the current research programme of finding empirically observable correlates of conscious states. This is because, according to Lakatos, research programmes are replaced only when a new research programme becomes capable of both explaining the success of its predecessor and predicting novel facts that its predecessor could not. While the first criterion can be met, as the explanatory success of correlative approaches can be attributed entirely to the fact that references to consciousness in successful hypotheses can invariably be replaced by references to awareness, attention, wakefulness, or responsiveness, it is not clear that the second criterion can. This is because the correlative approach contains a scientific element about the relationship between behavioural capacities and cognitive and physiological states

plus a speculative element about consciousness, and as such removing the element about consciousness does not directly entail the prediction of novel facts.

If we admit the above analysis, though, then we admit that consciousness plays no explanatory role even though it is a part of the current research programme. The number of facts explicable by the correlative approach to the study of consciousness is identical to the number of facts explicable by an approach that does not mention consciousness but simply mentions awareness and so on. This is because any fact that seems to be about consciousness in the current approach must be a claim about the relationship between some behavioural, cognitive, or physiological state and another empirically observable one, and it is thus simply the reference to the word “consciousness” or the phrase “conscious state” that must be dropped, not a set of empirically verifiable or falsifiable claims. Thus, consciousness only formally forms part of the current research programme, as a world in which it had not done so would still be able to empirically determine the truth of the same number of facts.

This means that, even under Lakatos’ model, the description of consciousness in scientific discourse is not guaranteed to have any implications regarding whether or not consciousness is indeed being studied or whether facts pertaining to consciousness are actually being discovered by that area of research. Even if consciousness should be possible to study scientifically, then, this would not prohibit reference to consciousness from entering into discourse that ultimately fails to study consciousness. As such, adopting Lakatos’ position does not avoid the paradox.

The problem then is not that Lakatos’ research programmes are a flawed way of looking at science but rather that they do not give us a means of defending the current endeavour to study consciousness scientifically. While these research programmes may not be required to justify the assumptions defended as part of the core of those areas of scientific investigation, if they can claim to understand something then they must form a part of that core in a way that is not purely formal. If the phenomena described by a research programme are real, then

introducing mention of them into that research programme should entail novel predictions. This is not the case for the correlative approach to a science of consciousness.

[5.6.2] Revisiting anarchism

Feyerabend has already been discussed in this thesis, but that discussion focussed on how the fact that a science must be able to empirically distinguish between claims is not incompatible with his anarchist project. The correlative approach meets this basic criterion, as it does allow for us to empirically distinguish between claims; to state that consciousness is correlated with reverberatory thalamocortical circuits is clearly based partially on empirically justified reasoning that would allow us to make that claim instead of the claim that consciousness is correlated with, for example, a certain kind of liver function. I have made a case against this reasoning already, but anarchism seems to give us an easy way to reject these reservations.

Anarchism rejects an overall philosophical picture of the functioning of science such that it is possible to state that science consists of a specific kind of activity or requires the use of specific kinds of claims. Feyerabend described the problem with assuming that a certain philosophical view on science is correct in a dialogue where the character named after Feyerabend responds to somebody claiming that the process of science was not well understood until Popper:

> I find this most surprising. Scientists have wrong views about the nature of science. Still, they make discoveries, initiate revolutions, constantly widen our horizon. Popper himself makes science a paradigm of knowledge. Popper, on the other hand, has the right view. Yet all we find in him are some silly and thoroughly uninformed suggestions about the interpretation of quantum

> mechanics – about the interpretation, mind you, not about the theory itself which was invented by muddleheads like Bohr, Heisenberg, Born, Schroedinger. What I infer from this paradox is that we must distinguish between the practice of science, which is complicated, not entirely transparent – but seems to get results and philosophical ideas which may be right, which may be wrong, but which have no influence whatsoever on that practice (Feyerabend 1991: 490–491).

Feyerabend's stance seems to have been to reject the notion that we can have an overall philosophical view that supersedes or is more epistemologically fundamental than our scientific views. Indeed, he stated that there could be no theory of reason because, "reason is constituted by actions which cannot be foreseen unless they are restricted by totalitarian measures" (ibid.: 518). He stated that in situations where individuals are unswayed by rational argument, "Even the most puritanical rationalist will then be forced to stop reasoning and to use *propaganda* and *coercion*" (Feyerabend 1975: 16).

It would be a misconception of Feyerabend's stance to state that he believed that we should allow the influences on science to be completely unconstrained. He stated that we should not allow unrestrained use of "objective, emotionally antiseptic knowledge" (Feyerabend 1991: 498). As such, it is precisely because philosophical views on the nature of science aim to constrain its possible influences that Feyerabend countered that it is actually these philosophical views that should be constrained. It does not seem to be the case that Feyerabend opposed philosophy because he considered it to be actually dangerous in such situations, but rather that it aimed to elevate itself above the study of science; he stated that he "cannot stand it when so-called thinkers not only presume to know things better than their fellow human beings… but put them on a lower level of existence" (ibid.: 496).

While his rejection of a singular picture of reality caused him to be accused of being sympathetic to relativism (Bhaskar 1975), Feyerabend's argument was not that there is no

objective reality, but rather that reality is best understood through a variety of approaches, and that these approaches will not fit neatly into a single methodological perspective (Feyerabend 1975: 21–23; Feyerabend 1991: 519–520). His stance was thus that attempts to lay out rules for what constitutes proper reasonable discourse on the nature of reality according to critical rationalist perspectives should be rejected (Feyerabend 1975: 154).

A reasonable question to ask here is, if the progress of science is enriched by it having varied surrounding influences then what exactly is supposed to be the problem with philosophy being one such influence? Even if philosophy does aim to be restrictive over science, Feyerabend seemed not to believe that it actually has any impact on its practice. As such, there does not seem to be any problem with having a pool of ideas developed separately by philosophers in their apparently irrelevant commentaries on the field as this simply adds to the totality of concepts that can be drawn from wherever they may be useful.

Anarchism, as defended by Feyerabend, was explicitly liberal in its description of which ideas about the nature of science are acceptable – Feyerabend even advised using self-inconsistent hypotheses in certain situations (ibid.: 74–75) – but it was also motivated by an implicit conservatism where the status quo of scientific ideas is defended against possible radical reinterpretation from philosophers. To demonstrate the inconsistency in this stance, I will compare it with the scientism defended by Ladyman and Ross. In their defence of scientism, they argued that “neo-scholastic” metaphysics should be rejected because it is either incompatible with or irrelevant to the approximately consensual scientific picture (Ladyman, Ross et al. 2007: 17–27), which is the only means we have of understanding reality (ibid.: 28). Their position was thus premised very heavily on a profound respect for science. The trouble is that bona fide institutional science, which they frequently deferred to as the jury when it comes to what should and should not be considered science, uses metaphysical concepts whenever it seems appropriate. Indeed, much of my criticism of recent attempts at a correlative consciousness science has been that it adopts such concepts inappropriately, with scientists referencing things like qualia and the self in extremely ambiguous interpretations

of scientific results that do nothing to further the work in their field. This leaves the authors in a predicament given that on the one hand that they believed science to be exclusively authoritative over the understanding of reality and that on the other hand scientists use concepts developed within the "neo-scholastic" metaphysics that we were supposed to have rejected.

Even if this is not a decisive argument against the scientism these authors defend, it demonstrates a tension in their arguments. Either "neo-scholastic metaphysics" should not exist and thus institutional science inappropriately uses metaphysical concepts, in which case anybody with an interest in science (including a metaphysician) who noticed such a thing would be able to refute claims made by these institutions, or institutional science is exclusively authoritative over the understanding of reality and thus the existence of "neoscholastic metaphysics" has been useful to this enterprise in providing the concepts it adopts. Either way, their argument against metaphysics is in tension with their argument in favour of scientism. They are supposed to be providing a defence of the authority of science and yet if it were to succeed and "neo-scholastic" metaphysics were to be annihilated completely, this would rob institutional science of concepts that it might otherwise have used, which would have made impossible parts of the current scientific worldview that the authors maintain metaphysics should exist to serve. The respect the authors seem to have then is not for the process by which our scientific worldview changes over time but rather specifically for the worldview defended by contemporary scientists. While the authors have claimed to be happy to have a philosophical perspective that will be rendered false given scientific change, if the philosophy of these authors became mainstream enough it would eradicate some of the conditions that would bring that change about.

Likewise, Feyerabend's anarchism was supposedly a perspective about liberating science from the oppression of imposing philosophical viewpoints, and yet attempts to provide such an understanding of science have been present during the most fruitful centuries of scientific development. We would have to understand it as being the case that such scientific

progress has taken place despite the presence of philosophical theories on the nature and limitations of science, and perhaps that scientific progress would be improved in the absence of such philosophical approaches. Yet, philosophical worldviews consisting significantly of many philosophers stating that some aspects of consciousness (such as qualia) lie fully outside the study of empirical science have contributed concepts related to consciousness that scientists readily accept and that form part of their discussions, which is just the area under question in this thesis. As such, if increasing the variety of concepts available to practitioners of science was Feyerabend's goal then restricting philosophical systems runs counter to this goal.

This response has targeted the consequences of anarchism rather than the truth of the anarchist's stated position, but this is because anarchism pointedly does not consist of a positive picture. Feyerabend stated that the philosophical position he developed in *Against Method* was "a joke" (ibid.: 489). There is some parallel here to Wittgenstein's ladder, where Wittgenstein used philosophical reasoning to liberate the reader from philosophical reasoning:

> My propositions are elucidatory in this way: he who understands me finally recognizes them as senseless; when he has climbed out through them, on them, over them. (He must so to speak throw away the ladder, after he has climbed up on it.) (Wittgenstein 1921: 6.54)

Indeed, the character named Feyerabend in his dialogue comments that he wrote his famous book following his in-depth reading of Wittgenstein's works (Feyerabend 1991: 489).

Instead of arguing against the anarchist's supposedly non-existent worldview, I will present two possible and opposing views about the relationship between philosophy and science that could be admitted by the anarchist, which are that either philosophical intervention into scientific practice is acceptable or that it is not.

If philosophical intervention is acceptable, then there is no reason why my arguments against the scientific discourse should be off limits. If it is not acceptable, then scientists should not have adopted the concept of consciousness in the first place. Either way, the attempt to produce a scientific correlative view of consciousness remains equally undefended.

The only way that the anarchist could defend the science of consciousness is by arguing that the sort of intervention from philosophy required to produce it is acceptable but that the sort of intervention from philosophy that I am offering here is not acceptable. Perhaps a case could be offered for this, although I am unable to derive it from my understanding of Feyerabend's position. Indeed, Feyerabend's strategy seemed to be to abandon philosophical concepts of mind that were unhelpful to science and so he might be thought to agree with me. As it happens, Feyerabend's eliminative materialism required certain commitments that I do not agree with, but this will form part of the discussion about modifying our conception of consciousness.

I find it unlikely that many anarchists would accept the possibility that only interventions of the kind that brought us our correlative approach to consciousness science are acceptable whereas those of the kind that I am offering are not, because this would imply that we can give clear demarcating criteria between that which is able to influence science and that which is not. While Feyerabend did not believe that everything should be allowed to influence science, so long as my suggestions are not restrictive of science I do not see why they would not be permitted. Additionally, given that my suggested modifications to the currently existing approach would not rule out any scientific procedure, there is no reason to see them as inappropriate.

[5.6.3] Summary of attempts

It is fair to say that neither Lakatos' nor Feyerabend's work requires us to reject the view that the correlative approach genuinely studies consciousness, but this has not given us any reason to believe that it actually does. Indeed, the fact that they were both so liberal in terms of the sort of claims they allowed to constitute a scientific theory entails that it does not even matter if a discipline claims to study certain phenomena even if such claims are false. If a false claim was to form the core of a scientific study, as long as it resulted in auxiliary hypotheses that are empirically testable it would be acceptable under Lakatos' perspective. If a false claim was to be central to a scientific study, as long as it increased the variety of theories a scientist could have and observations she could make, thus enriching the diversity of human understanding, it would be acceptable under Feyerabend's perspective. If, then, a claim is false but also does neither of these things then there is no reason for the anarchist or the defender of research programmes to favour this approach.

The question we should ask now is whether claims central to the science of consciousness are true – and this will depend upon our understanding of the nature of consciousness.

If we wish to understand the nature of consciousness and its relationship to the rest of our reality, it is in our interest to deny the coveted claim of "science of consciousness" to the current attempt to provide a correlates-based science of consciousness because this denial provides a strong motivation to critically analyse any perspective that claims to use empirical knowledge to discover facts about consciousness. Our critical analysis of recent attempts to study consciousness science may lead us to believe that consciousness cannot be studied scientifically, but, for methodological pluralists like Lakatos and Feyerabend, philosophical justification of the claims we make about the nature of phenomena is a wholly different matter to whether they form the core of a scientific methodology. As such, any reason we have to disbelieve in the truth of certain claims about consciousness – such as that we cannot

empirically distinguish between its presence and absence – does not prevent there being scientific theories that rely upon the falsehood of that claim.

If we wish to be able to justify the claim that consciousness is being studied by such a science though, these responses are inadequate. Lakatos and Feyerabend specifically give provision for scientists to make certain claims that are irrelevant to their study or even untrue. Whatever we are to make of this, if it is no argument against current scientific approaches to the effect that they fail to study consciousness, it is equally no argument that consciousness should not feature in a scientific study as a proper object, where true and false claims about it can be ascertained on the basis of observation. As such, since consciousness can both be thought of as something that science should be able to study and something that science cannot study, the paradox persists.

[5.7] An illustrative example of scientific development from quantum mechanics

Simply switching from one conception of science to another offers us no guarantee of a resolution to the paradox, which is because these perspectives tend to offer accounts of science that differently account for the relationship between observations and theory, whereas this has no bearing on the fact that observations have no impact on the truth or falsity of claims about consciousness.

While it is possible to modify our conception of science such that it is not required to understand the nature of consciousness, this does nothing to remove the Consciousness Science Paradox. Even if science is not required to be able to provide a model of consciousness justified by empirical observations, this does not entail that consciousness should not be possible to comprehend in scientific terms.

What constitutes proper scientific practice is disputable and indeed changes over time. Yet, the fact that there is no known accurate description of science should not lead us to the

conclusion that scientific standards are entirely arbitrary and that all interpretations of phenomena, even if given by respected scientists, are equally valuable. There is a crucial distinction between interpretations of the observations made by scientists that go on to form a major part of scientific understanding and practice and interpretations that fail to do so. Indeed, even though scientific results can sometimes overthrow our preconceptions about the world and how we can observe that world, not all interpretations are equally able to perform this function. This can be demonstrated using an example taken from quantum mechanics.

There is a well-known experiment in quantum mechanics known as the "dual-slit experiment," where photons or electrons are fired toward a material obstructed by a surface with two parallel slits through which the light or electrons can pass. The experiment was performed with light (Young 1804) and then over a century later with electrons (Davisson & Germer 1927). An initial unusual result was found that rather than two bands of particles being found on the material, several bands of particles were produced. This pattern is known as an interference pattern and is consistent with a wave flowing through the slits, producing two separate waves from each slit at which there are certain points where the waves intersect, with the particles hitting areas on the material where the peaks of such a wave line up much more frequently than where a peak meets a trough. The result was surprising because it is inconsistent with particles (conceived as solid objects hurtling along a given trajectory) each passing through one of the slits, where we would expect them to collectively produce two bands. Indeed, even if a single particle is fired toward the material at a time, the interference pattern is still gradually produced over time, suggesting that the behaviour of each individual particle is consistent with that of a wave flowing through both slits simultaneously (Cox & Forshaw 2012: 20–24).

Since these results, further experiments have been conducted with stranger implications. The effect of measuring which slit the particle goes through has been shown to be to reduce the interference pattern, so that a particle behaves less like a wave if its path is measured. The more accurately the path of the particle can be derived from measurement,

the less the particle behaves like a wave (Wooters & Zuerk 1979). This cessation of wavelike behaviour is described as the collapse of the wave-function of a particle. In an even stranger result, if a detector measuring the path of the particle is present that subsequently erases the "which-path" information, the interference pattern is restored (Walborn et al. 2002). As such, the mere presence of a detector does not interfere with the wave-function of a particle; the particle's wave-function only collapses if information about it is gathered and *retained*.

What this means in terms of scientific observation is that the world is not quite how we might have imagined it before; there is not a single objective state of affairs that we observe through measurement, but rather the means by virtue of which we gather information about the world partially determine the state of the observed phenomena, such as by collapsing wave-functions. Although we can still derive facts about the state of the world through empirical observation, we now have to incorporate the effect of our measurements into that understanding and subsequently our understanding of empirical observation itself has to be, to some extent, revised.

Now, there are multiple ways we can revise this understanding. There are two broad views that I wish to consider here to demonstrate the distinction between such a revision that naturally follows from the continuation of the research projects involved in the study of quantum mechanics and one that does not. We can interpret the results of the dual-slit experiment by claiming that the wave-function is collapsed by there being some persisting physical effect that has resulted from the particle passing through a single slit as opposed to the other, or we could interpret the results of the experiment by claiming that the wavefunction is collapsed by subjective perception (Von Neumann 1932: 417–445).

The former of these interpretations seems amenable to empirical investigation in an uncontroversial sense. Even though the end result is a modification to our understanding of what empirical observation must involve (i.e. that it must involve collapsing wave-functions), the distinction between empirical evidence that would support or contradict such a claim is quite clear. If the wave-function of a particle is preserved, producing an interference pattern,

only when we can find no physical effect that is specific to it taking one path rather than the other (meaning that the physical effect would, as far as we can tell, be the same if it had taken the other path), this would provide evidence for such a claim. If we could consistently find physical effects that would tell us which slit a particle had passed through even while the wave-function is preserved, this would provide evidence that such a claim is false.

The latter interpretation, where wave-function collapse is produced by the subjective perception of a certain state of the world, is not amenable to scientific investigation for a straightforward reason. Von Neumann stated that quantum mechanics describes the state that the world is in when it is not observed because describing the nature of the world as it is observed requires us to describe additional processes (ibid.: 420). In the case of wavefunction collapse, we might argue alongside Von Neumann that to describe the way a particle behaves when its path is not observed is a separate task to describing the way the particle behaves when its path is observed, since it will behave like a wave in the former case and a particle in the latter case. This does not directly follow from the previous interpretation; it may be possible to make an observation based on the effects produced by the particles in certain environments but it does not follow that it is the observation itself that produces these effects. This is simply one possible interpretation of the causes of wave-function collapse. The problem is that, while we can in principle empirically distinguish between cases where the physical effects are present and where they are absent, we cannot distinguish between cases where the physical effects are present but we do not in any way observe that they are and where the physical effects are present and we do observe that they are. This comparison is what would be required in order to ascertain whether or not the observation itself is responsible for wave-function collapse. We would have to somehow compare an observation we have not made with an observation we have made, which is of course impossible given that both states of affairs would need to be observed in order to make any comparison. The two cases are thus empirically indistinguishable.

I am not sure which philosophical view on the nature of science would be amenable to the observation-dependence interpretation of wave-function collapse. It is not falsifiable, it does not lead to any new hypotheses that can be verified or falsified, and there can be no future investigation that will make it clearer what kind of relationship exists between observed and unobserved phenomena since we will always be limited to the study of the former. The only perspective I can think of that would find such an interpretation acceptable would be institutionalism provided this is a widespread opinion accepted within institutional science. But we have already found reasons to reject institutionalism.

The claims made within quantum mechanics have been described as particularly unclear in terms of what the relationship is between the reality it describes and observation (Weyl 1949: 264). In general, there is a measurement problem; it seems difficult to account for the fact that measurements yield well-defined outcomes despite the fact that particles do not have well-defined positions if those positions are left unmeasured (Allahverdyan, Balian & Nieuwenhuizen 2012: 6). Like in any area of science, such as the examples shown in neuroscience and cognitive science above, we may attempt to plug gaps in our understanding with theories that entail no further empirical investigation, and the theory that the world exists a certain way outside of subjective perception that is distinct to how it appears inside subjective perception is exactly such a theory. The trouble is that such claims are antithetical to the development of empirical science; they provide us with no possible means to compare the two situations.

The progress of quantum mechanics has modified our understanding of the role of observation in interpreting the state of the world, but this does not entail that the role of observation is flexible in all directions. Given that the observation-dependence interpretation of wave-function collapse is that the world exists differently inside and outside of observation, changing our interpretation of what constitutes an observation would not make a difference anyway. This is related to the issue with the correlative approach to a consciousness science; since certain assumptions have to be made prior to any observation, such as that conscious

states only exist where the capacity to produce particular behaviours exists, there is no way to modify our conception of what constitutes observation to allow for these assumptions to be supported by observation.

We might counter that scientific theories are often rejected or accepted in terms of other features such as elegance or simplicity (Feyerabend 1975: 26) and thus that a theory does not need to be more amenable to empirical verification or falsification than another in order to replace it. Even if this is so, however, there is no need for us to give strict demarcating criteria here between scientific and non-scientific theories, as I stated when discussing what it should mean for something to be amenable to scientific study. There is only a need to acknowledge that a scientific perspective must be able to distinguish between states falling within its subject matter on the basis of observation.

If the distinction between two scientific theories could be made purely on the basis of elegance this would mean that they are both equally capable of distinguishing between claims on the basis of observation and thus the fact that one was replaced by another does not contradict any of my claims made so far. With the subjective perception interpretation of quantum mechanics above on the other hand, this is an explanation that adds no capacity to empirically distinguish between claims, given that it has no empirical consequences, but if it became widely accepted it would *remove* some capacity to empirically distinguish between claims since it precludes a solution to the quantum measurement problem that can be empirically tested. The same applies to the correlative approach to consciousness science; it adds no empirical content but by widely accepting it as a science of consciousness we are supporting the preclusion of a study of consciousness that does add empirical content.

As such, when I state that the conception of science naturally changes given scientific progress, I refer to cases such as the interpretation of wave-function collapse that is amenable to empirical confirmation or refutation as opposed to cases such as the interpretation that places the cause of wave-function collapse outside empirical investigation. Although science is host to a wide variety of interpretations both by practitioners of science and those interested

in its philosophical foundations, only some of these interpretations contribute to the explanatory power of science.

[5.8] Summary

We began this chapter by establishing the two claims of the paradox, that PC states both must be identifiable through empirical means and that they cannot be [5.1]. We then inferred that "consciousness scientists" are actually studying certain behavioural tendencies and cognitive capacities rather than consciousness [5.2]. We considered the paradox to be a philosophical problem [5.3] and assessed some attempts to re-evaluate our conception of science [5.4]. First, we assessed the possibility that science could be conceived in a way that would allow us to claim that it is able to explain consciousness, but after examining Popper's falsificationism [5.5.1], Kuhn's view on scientific revolutions [5.5.2] and the Bayesian approach [5.5.3], we found that none of these views gave us an account whereby consciousness could be said to be explicable. We then assessed the possibility that science could be conceived in a way that would allow us to claim that it does not need to explain consciousness [5.6], as in the cases of Lakatos' perspective on research programmes [5.6.1] and Feyerabend's anarchism [5.6.2], only to find that neither of these perspectives offered us a means of avoiding the paradoxical conclusion regarding consciousness. We followed this by exploring the possibility that future scientific developments could change our understanding of observation such that consciousness may become explicable, but showed using an example taken from quantum mechanics that the concept of observation cannot be arbitrarily modified and still yield a study that can differentiate between possibilities on the basis of observable phenomena [5.7].

The survey of positions given here should have demonstrated that the Consciousness Science Paradox is not caused by a limitation in our understanding of the nature of science.

This leaves open the only other avenue for a solution to the problem, which is to evaluate the relationship between consciousness and the world.

6

Evaluating the Nature of Consciousness

We are left with the same incompatible claims of the Consciousness Science Paradox if we adopt any one of the several conceptions of science examined above. It must be true both that PC states should be possible to study empirically, and that it is impossible for PC states to be studied empirically. If we cannot avoid this conclusion by abandoning a particular perspective on the nature of science, we might try to avoid it by abandoning a particular perspective on the nature of consciousness. We might reason that it is a defect of our understanding of consciousness given that it carries with it the promise that consciousness should be able to become an ordinary aspect of scientific understanding given suitable observation and experimentation, yet deprives us of any tools by virtue of which such an understanding could be made possible. Our only recourse, we may suspect, is to abandon our preconceptions about consciousness and adopt a different metaphysical perspective.

This is a far cry from a sure-fire means of solving the problem. There are wildly differing perspectives on what consciousness is, as stated at the outset of this thesis, but the problem cannot be isolated to a specific metaphysical perspective. In fact, there are multiple major positions in the metaphysics of consciousness that run into the same paradoxical result.

I will demonstrate how the problem persists in relation to three metaphysical perspectives on the nature of consciousness: **Physicalism, Property Dualism** and **Russellian Monism**.

[6.1] Physicalism

The idea that consciousness is nothing other than the physical is, broadly speaking, a physicalist idea. Physicalism, a term often used interchangeably with materialism, is the thesis that everything is physical, or that everything is at least realised by physical states. Neurath coined the term to express the goal of forging a unified science unburdened by philosophical speculation and able to give a system of laws whereby "every phenomenon is tested by means of sound, light, etc., but sound and light play no part in the final scientific presentation" (Neurath 1931: 620). Instead, everything would be explainable purely in the logical-mathematical language of physics. He then introduced the term to describe the physical language that the project aimed to describe reality using.

> In a sense unified physics is physics in its largest aspect, a tissue of laws expressing space-time linkages–let us call it: *Physicalism*. (ibid.: 620)

The logical positivism that characterised the philosophy of the Vienna Circle has not been a widely endorsed philosophical approach at least since Quine's attacks against dominant empiricist views in 1951. Quine argued that for all its ostensive commitment to purely empirical facts, empiricism relied upon the "unempirical dogma" that there is a boundary between analytic and synthetic statements (Quine 1951: 34) as well as the assumption that it is possible that a statement, "taken in isolation from its fellows," can be confirmed by sense experience, two dogmas which are "at root identical" (ibid.: 38).

Following the diminishing influence of logical positivism and discontentment with the behaviourism espoused by Ryle (1949), the view that the mind needed to be taken seriously came back into mainstream philosophical discussion, with reductive physicalist approaches having a prominent position in this debate.

The usage of the term “physicalism” has shifted since its introduction, such that philosophers now claiming to be physicalists are very unlikely to be committed to a programme of direct translation between experiential reports and physical science. The only commitment that has remained is that physical science captures the essence of fundamental reality.

If we were not to defer to the physical sciences in our description of the physical, it would not be clear what view the word “physicalism” refers to at all. For instance, Stroud suggested that we could describe the physical world not as “the world described in the terms of the physical sciences,” but rather as “the world that the physical sciences describe” (Stroud 2000: 65), and as such we could define physicalism as the view that nothing exists other than that constituting the world, which the physical sciences aim to describe. This would fail to exclude even subjective idealism as a physicalist position however because, for subjective idealism to be true, it is simply the case that the world that physics aims to describe is an immaterial world made up of mental content. If this is so, physicalism could only really be described in contrast to dualism; provided we believe that consciousness is composed of the same stuff that the world physics aims to describe is composed of, we would be physicalists. Yet, this is characteristic of monism generally, which includes hugely varied positions with radically different interpretations about the nature (and indeed the existence) of matter, and so if physicalism can denote a more specific perspective, it must be possible to contrast it with other forms of monism. Consequently, it seems that for the word “physicalism” to have any theoretical use, it must denote a worldview in line with the former of Stroud’s definitions; it is a thesis about describing all of the phenomena in the world in the terms of the physical sciences.

Physicalism can broadly be broken down into two kinds. Type-physicalism, which claims that types of mental state are identical with types of physical state, such as pain being c-fibres firing, and token-physicalism, which holds that each token of a mental state, such as an instance of pain, may be identified with a token of a physical state, such as an instance of c-fibres firing, but that pain as a type of mental state may be realised by a multitude of different types of physical state. I will address each of these physicalist perspectives in turn.

[6.1.1] Type-physicalism

Type physicalism is the physicalist perspective that holds that mental states are reducible to physical states. We will be concerned with this reductive strategy here, although, as Tartaglia observed, Smart and Place, two paradigmatic examples of reductive physicalists, rejected phenomenal qualities (Tartaglia 2013: 819–820) and so they would have fit comfortably into the ranks of eliminativists. All that is required for Smart's metaphysics of the mind is that the concepts we use to denote sensations are simply those that are applicable under particular circumstances, and not because of the presence of a particular subjective class of properties.

Whether type-identities are compatible with the seeming possibility that the same mental states could be realised by different physical states is disputable, with some arguing that it can be (Jackson, Pargetter and Prior 1982), and others that it cannot be (Fodor 1974). I will address the former possibility here and the latter possibility in the next section [6.1.2].

A popular approach for type-identity theorists has been to adopt what has been called the Phenomenal Concepts Strategy (Stoljar 2005). This strategy most typically consists of arguing that phenomenal concepts are particular kinds of "recognitional concepts" (Loar 1990: 87). This is argued to allow for a physical property to "trigger" the concept of being in a certain state and thus for the phenomenal property referred to by that concept to refer to that physical state, even while the identification between the phenomenal property and physical state can only be made *a posteriori* (ibid.: 88). This strategy has been used by many physicalists since the 1990s wishing to describe the relationship between consciousness and physical processes without denying the existence of intrinsic qualities of consciousness (Tartaglia 2013: 2) and is not only an option for those wishing to endorse reductive physicalism, but also those who wish

to identify phenomenal properties with functional properties that can be physically realised (Lycan 1987).

The phenomenal concepts strategy relies upon "the cognitive disparateness of perception- and introspection-based concepts" (Tartaglia 2013: 11), thus allowing both concepts to refer to the same thing without us being able to recognise this identity. The success of this strategy has been measured in relation to its capacity to deal with Kripke's argument that the conceivability of conscious states existing in the absence of any specific physical state entails that the two cannot be type-identical (Kripke 1972: 144–155), with some arguing that it fails at this benchmark (Stoljar 2005: 485–486). Tartaglia argued that the adoption of phenomenal concepts forces the physicalist to adopt the perspective that we misrepresent physical states as phenomenal (Tartaglia 2013: 10–15), in which case they must accept that phenomenal properties are "mythological" (ibid.: 16–17). If this is so, the reductive position is really an eliminative position (which we will discuss in [7.3]).

Even if reduction does not collapse into eliminativism and can avoid the above charges, the Phenomenal Concepts Strategy nonetheless inevitably falls victim to the Consciousness Science Paradox. I cannot know *a priori* that pain is identical to c-fibres firing. I must then be able to correlate my phenomenal state with my physical circumstances, such as my neurophysiological state. Yet, this runs into familiar difficulties. To determine, for instance, that a certain neurophysiological state must correlate with the conscious state of feeling happiness, I must be able to determine when that conscious state is present and when it is absent, which we have already established is impossible. Furthermore, I would need to overcome the difficulty of determining whether states of both kinds have occurred simultaneously, which we have also already established is impossible.

Specifying that phenomenal properties exist does nothing to save us from the Consciousness Science Paradox. The paradox is produced by two claims, neither of which a type-physicalist position allows us to deny. On the one hand consciousness should be amenable to empirical investigation, which type-physicalism guarantees by claiming that any

conscious state is a particular physical state, and on the other hand consciousness cannot be amenable to empirical investigation, which is guaranteed by the fact that no empirical correlation between phenomenal qualities and physical states is possible.

[6.1.2] Token-physicalism

Token physicalists accept that there is nothing more to any mental state than being in a particular physical state but deny that this is because types of mental state are identical to types of physical state. There are two ways that this can be accepted.

The first is that mental states follow laws that are not reducible to physical laws. This sort of perspective was held by Fodor, who argued that laws determining which mental state follows from which other mental state under certain circumstances are not laws that can be reduced to laws about which brain state follows which other brain state under certain circumstances (Fodor 1974). I will give one of the arguments that Fodor used for this position: if every psychological type can be reduced to a number of distinct physical types, we might state that psychological state P1 is reducible to either physical state S1, S2 or S3, whereas psychological state P2 is reducible to either physical state S4, S5 or S6, and it could be true that S1 always leads to S4, S2 always leads to S5 and S3 always leads to S6. If, then, it is a law of psychology that P2 follows from P1, then we might attempt to claim that this law can be reduced to the physical law that either S1 leads to S4, S2 leads to S5, or S3 leads to S6. Fodor argued that this would be a mistake; laws are not of the form “either X happens, Y happens or Z happens.” He gave an example:

> I think, for example, that it is a law that the irradiation of green plants by sunlight causes carbohydrate synthesis, and I think that it is a law that friction causes heat, but I do not that it is a law that (either the irradiation of green

> plants by sunlight or friction) causes (either carbohydrate synthesis or heat). Correspondingly, I doubt that 'is either carbohydrate synthesis or heat' is plausibly taken to be a natural kind predicate. (ibid.: 109)

The second way we can accept token-identities while denying type-identities is to state that mental states follow no strict laws at all, even though all mental states supervene on physical states. Supervenience is a one-way relation whereby "there cannot be two events alike in all physical respects but differing in some mental," but there is no restriction on there being two mental events that are alike but differ in some physical respects (Davidson 1970: 214). This sort of perspective was held by Davidson, who argued for anomalous monism, a position that is often construed as physicalist, and is certainly compatible with physicalism in that all mental tokens are physical tokens, without it being the case that there are any laws determining which mental tokens should be identical to which physical tokens.

While both Fodor and Davidson defended these positions as regards mental states or events more generally, with Davidson's interest being in events with intentional content, our interest is in whether either form of token identity theory would provide shelter from the Consciousness Science Paradox, and so we will need to consider the extent to which either apply to the relationship between conscious states and physical states.

In Fodor's case, his account of the discrepancy between psychological laws and neuroscientific laws would avoid the paradox only if it could be established that the empirical science of psychology is able to give an account of consciousness that avoids the paradox. Yet, the paradox as outlined in this thesis already applied to any position that attempts to describe correlates of consciousness, whether these are neurophysiological states or cognitive states. In fact, one of the key examples of a flawed perspective highlighted was that of Bernard Baars [4.6], who gave his description of consciousness in terms of certain computational and information-processing capacities, but still relied upon (indeed pioneered) a problematic contrastive approach to achieve this goal, whereby the presence of conscious states as

evidenced by the presence of recognisable conscious behaviour was contrasted with the absence of those conscious states as evidenced by the absence of such behaviour. I refer the reader back to chapters 2 and 3 where I outlined in detail the problem with this approach.

To bring the point back around to the current matter, if the field of psychology is in no way more able to empirically identify correlates of consciousness than the field of neuroscience is, whether these are behavioural, cognitive or information processing capacities, then whether or not psychology can be reduced to neurophysiology the paradoxical result will be the same.

Regarding a Davidsonian approach, it is first important to state that Davidson would likely not have held this approach with regard to phenomenal qualities, as he rejected the idea of there being "such epistemological intermediaries as sense data, qualia, or raw feels" (Davidson 1988: 52).

Our concern here, then, is whether an anomalous monist position has anything to say that would allay our concerns regarding the consciousness science paradox or offer us a way of avoiding it. In some senses, the position outlined in this thesis agrees with what has been argued for by Davidson. Davidson argued for "the Principle of Causal Interaction," which states that "at least some mental events interact causally with physical events" (Davidson 1970: 208). This is very much in line with the earlier conclusion in this thesis that conscious states must be causally efficacious. Similarly, Davidson's conclusion that "there are no strict laws at all on the basis of which we can predict and explain mental phenomena" (ibid.: 224) is in line with the conclusion arrived at in this thesis that we have no means of determining whether a conscious state is present in many situations on the basis of empirical observation. These points are only superficially similar though; the sense in which Davidson means that mental events are anomalous is not the same sense we might say, as a result of the considerations found in this thesis, that conscious states are anomalous.

For Davidson, mental events are anomalous because the manner in which we attribute mental events is different to the manner in which we attribute physical events such that no lawlike relation could exist between the two. He argued that we attribute a particular mental

event to somebody by assessing how that mental event fits into an overall pattern of mental events such that “the content of a propositional attitude derives from its place in the pattern” (of beliefs, preferences, hopes, fears, expectations and so on) (ibid.: 221), whereas “it is a feature of physical reality that physical change can be explained by laws that connect it with other changes and conditions physically described” (ibid.: 222). For Davidson, it is because the non-strict laws determining whether such-and-such mental event is present are of a different kind to the strict laws determining whether such-and-such physical event is present, such that if we are to remain faithful to either “proper source of evidence” there cannot be “tight connections between the realms” (ibid.: 222).

As such, the anomalousness to which Davidson referred is not the same anomalousness to which I refer. Indeed, Davidson argued that we must suppose that the mental event happening within a person exists within an overall pattern of mental events if we are not to “forego the chance of treating them as persons” (ibid.: 222). This is where our positions part; the critique offered in this thesis has been that even in situations where an individual would ordinarily be supposed to have no capacity for reason, such as when they are under general anaesthetic, we still have no means of excluding the possibility that conscious states are present. Regarding Davidson’s position and his focus on mental events rather than conscious states, it is clear then that his subject matter is simply different to that of this thesis.

The difference in our positions can be primarily attributed to the fact that Davidson was primarily concerned with the content of propositional attitudes here and not conscious states. Indeed, if we ignored Davidson’s intended conclusions and co-opted his reasoning regarding mental and physical events to instead be about conscious and physical states, this reasoning would run into difficulties. Davidson drew temporal relations between mental and physical events that have been argued in this thesis would be impossible to draw between conscious states and physical sates. For instance, anomalous monism relies upon here being events that can be given both a mental and physical description (ibid.: 224). Whether or not a conscious state is even simultaneous with another observable event, though, is something we

have already found is not possible to determine. Since mental tokens are, under anomalous monism, identical to physical tokens, this point would require us to know when conscious states occur in order to establish if they do indeed occur alongside any other physical event, if we were to extend Davidson's thesis to be about conscious states and avoid the paradox. Let us take an example that Davidson gave in his argument against Brentano's thesis that the mark of the mental is that it exhibits intentionality. Davidson stated:

> Take some event one would intuitively accept as physical, let's say the collision of two stars in distant space. There must be a purely physical predicate '*Px*' true of this collision, and of others, but true of only this one at the time it occurred. This particular time, though, may be pinpointed as the same time that Jones notices that a pencil starts to roll across his desk. The distant stellar collision is thus *the* event *x* such that *Px* and *x* is simultaneous with Jones's noticing that a pencil starts to roll across his desk. (ibid.: 211)

With this line of reasoning being used to discuss conscious states we have a problem. Jones being in the conscious state of perceiving his pencil rolling across the desk, for instance, could not be pinpointed as occurring at the same time that two stars collided. As such, both the justification for anomalous monism and the proposition that mental events are identical to physical events would, if extended to be about conscious states, not be possible to demonstrate for reasons already outlined in this thesis.

Finally, it is simple to demonstrate that the supervenience relation that Davidson described with regard to mental events and physical events could not be identified in the case of conscious states. Supervenience is a relation whereby it is possible for a physical event to vary without a mental event varying but it is not possible for a mental event to vary without a physical event varying. If this relation were to hold between a particular conscious state and physical state, this could only ever be identified if we first established that these two states

were at least present at the same time, and, for the same reasons as those just discussed with relation to anomalous monism, this cannot be established.

Token physicalism, in short, does not provide us with a means of avoiding the paradox. For physical tokens to be identical to conscious tokens, it would have to be the case that the two are simultaneous in time. Since the paradox arises prior to establishing that this is the case for any pair of such tokens it follows that these positions only really get off the ground if the paradox is set to one side.

[6.1.3] A general conclusion about physicalism

The damage of this paradox can be felt by all positions that describe conscious states as being physically realisable. This includes reductive physicalism, any form of functionalism that accepts type-identities between physical and mental states, and any form of panpsychism for which brain states have phenomenal properties. For an example of such a panpsychist perspective, Galen Strawson argued that he was "happy to say, along with many other physicalists, that experience is 'really just neurons firing'" (Strawson 2006: 7) but that "physicalism… entails panexperientialism or panpsychism" (ibid.: 25). As such, Strawson's perspective is a clear example of a theory of the physical realisability of conscious states.

The problem cannot be avoided either by denying that consciousness plays a significant role in cognition even while it is physical, as the physical epiphenomenalism described by Flanagan suggested (1993: 130–131). To recap from our earlier discussion of epiphenomenalism, this perspective is that conscious states are physical processes that do "nothing useful," such as the brain going into a "funny oscillatory state that lasts a few seconds" (ibid.: 130). The trouble is that, for this view to be true, such a funny oscillatory state must be able to result in us making true claims about our conscious states and thus the conscious state must still play a role in our capacity to report those states. Yet, if this is the case then the

problem is just as bad if not worse; we cannot empirically identify the useless physical state that make our claims about consciousness true because we cannot empirically distinguish between the presence and absence of a conscious state. Denying that the physical state plays a useful role in cognition simply reinforces the fact that a conscious state could be present in any given physical state, reasserting the already significant difficulties in finding a method of distinguishing between the presence and absence of conscious states.

Physicalism thus offers no means of avoiding the Consciousness Science Paradox. It equally does not matter if we adopt a non-reductive position under which there are, for instance, functional states that could also be realised by non-physical states (Putnam 1967: 436). It is simply the fact that such positions defend the stance that there are physically realisable instances of phenomenal states that results in them falling victim to the paradox.

[6.2] Property dualism

If we run into the paradox when we assume that consciousness is physical, we might seek to avoid it by assuming that consciousness is not. For property dualists, conscious states are distinguished from physical states on the basis of the private, ineffable, qualitative nature of phenomenal states. Nagel argued that the only way to know "what it is like" to be in a conscious state is to be in it; there is no way to know through third-person observation or from description alone how it feels to undergo a particular conscious experience (Nagel 1974). Jackson argued that knowing all the physical facts does not entail knowing all the mental facts about a situation, such as in his famous Mary in the black and white room example (Jackson 1982). Chalmers argued that if we knew exactly how our bodies worked, including how we produce the behaviour we do, such as flinching away from hot things that touch our skin, we would still not know why any of these states are accompanied by conscious states (Chalmers 1995: 5). For the property dualist, it is not that phenomenal properties are physical properties

that we conceive of in a different way, but rather that they are properties that are numerically distinct from physical properties. As Chalmers put it, “after fixing all the physical truths, God had to do more work to fix all the truths about consciousness” (Chalmers 2001: 124).

Chalmers’ position was that phenomenal facts are not entailed by physical facts and thus that materialism is false (Chalmers 1996: 123). He argued this point by first stating that all physical facts are fixed by microphysical facts (ibid.: 77) but that this fixing does not apply to consciousness. To demonstrate this, he defended the conceivability of phenomenal zombies (ibid.: 94–99), the conceivability of inverted spectra (ibid.: 99–101), and conclusions drawn from the knowledge argument (ibid.: 103–104). More recently, Goff has defended the view that it is the conceivability of ghosts, beings who are subjects of experience without bodies, that are the more plausible threat to physicalism (Goff 2010), but the broad argument is the same; in positing one set of facts (i.e. the physical or phenomenal) there is nothing compelling us to posit the other.

Although Chalmers’ position distinguishes phenomenal qualities from other physical properties, we are not required to posit a separate non-physical substance to account for these, and as such “consciousness supervenes *naturally* on the physical” (Chalmers 1996: 124). It is thus simple to see why Chalmers has been such a vocal advocate for correlative approaches to understanding consciousness; if there is a distinction between physical and non-physical properties, we will need to see the circumstances under which the two occur together in order to determine what the conditions are for us to be in a particular conscious state. Chalmers championed the need to correlate first-person and third-person data (Chalmers 2003: 1111) and lauded neuroscience for the “major role” it “undoubtedly” has to play in our understanding of consciousness (Chalmers 1998: 219). We have no need to distinguish between the theory that non-physical states correspond with physical types or tokens here because, as shown in the previous section, we should expect in either case for there to be some identifiable correlation between conscious states and the physical.

Since we have already established that epiphenomenalism is not a viable position, we could present a viable argument against property dualism by defending the view that there is no way that physical and non-physical phenomenal facts can follow from one another, but this is not our concern here. Our concern is whether property dualism would enable us to avoid the Consciousness Science Paradox. The strength of a dualist position here could be found in its conceptual separation of consciousness and physical states such that we are not forced to suppose that the two have a strong relationship at all, thus undermining the premise that we should be able to identify which state corresponds with which other. This is not a strength that seems to have been taken advantage of at any point, as dualism seems to have had a strong reputation over the centuries for presenting arguments that assert the correlation between specific mental states and specific physiological states. Descartes accepted such a view based on the science of his time:

> My next observation is that the mind is not immediately affected by all parts of the body, but only by the brain, or perhaps just by one small part of the brain, namely the part which is said to contain the 'common' sense. Every time this part of the brain is in a given state, it presents the same signals to the mind, even though the other parts of the body may be in a different condition at the time. This is established by countless observations, which there is no need to review here. (Descartes 1641: 59–60)

This perspective then, already criticised in some detail, that conscious states must be related to specific neurophysiological types or tokens, and moreover that this relationship can be established by empirical observation, has developed naturally from views held about the mind-body relationship as long ago as Descartes' time. Tallis has traced the view that the brain is the seat of the soul, mind or consciousness to before Hippocrates (Tallis 2011: 29).

This same basic view is reflected in Chalmers' claim that progress on our understanding of consciousness can be made by drawing up correlations between conscious states and neurophysiological states (Chalmers 1998). It is easy to see why the relationship between phenomenal and physical properties is assumed in property dualism; the position relies on there being a distinction between the two, and yet we can only know the truth of dualism if we can gain knowledge from both the physical and non-physical domain, and so they must both be supposed to be distinct and yet for both to be able to influence the state of our knowledge at any particular time. The tendency for some dualists to end up defending epiphenomenalism is also understandable given that the distinction between phenomenal and physical properties is often drawn between the former being intrinsic and qualitative and the latter being extrinsic and causal-structural.

Epiphenomenalism is not a viable position, as discussed near the beginning of this thesis, since conscious states, if causally inefficacious, should not be able to influence our capacity to make true claims about them. The only option available to the property dualist, therefore, is for it to be the case that phenomenal properties do have an effect on our behaviour, including the words that come out of our mouths. Given that this must be so if property dualism is true, we should be able to make empirical observations of the causal powers of phenomenal properties, and yet we cannot. The reason we should be able to make such empirical observations is that, if phenomenal properties have particular causal effects, we will be able to identify the presence of phenomenal properties by the presence of such effects. There are many reasons why we cannot, including the fact that we are unable to determine whether a phenomenal property is absent even where there are no familiar behaviours associated with that property unless we make *a priori* assumptions about what the total causal footprint of that phenomenal property is. This is to become embroiled straightforwardly in the Consciousness Science Paradox, and thus this is a major problem for property dualism.

Chalmers seemed to defer to epiphenomenalism in some form on certain occasions. He argued that the causal efficacy of non-physical properties "raises more problems than it solves" in that it relies on the assumption that future discoveries in physics will not be able to describe the causes of physical events in purely physical terms, which seems unlikely (Chalmers 1996: 156). He stated that it also assumes that there will be a gap in cognitive science that can only be filled by the introduction of non-physical properties, which equally seems unlikely (ibid.: 156). What may appear to be a more promising avenue for his philosophical position is his deference to a Russellian monist perspective, which I will discuss presently.

As a brief note, I would just like to point out that substance dualism would be subject to the same problem. Whether we maintain that non-physical substances interact with the body via a certain locus in the brain or whether God initiates phenomenal states given certain physical conditions, a precondition for supposing that this sort of model accurately describes reality is the view that both kinds of state occur together, or one following another, and, given that empirical observation cannot reveal this, any speculation about how the two correspond is at best baseless and at worst requires the acceptance of the two contradictory claims of the Consciousness Science Paradox. In the case of property dualism, it is presumably only possible to find correlates of consciousness as opposed to perceiving the properties of consciousness themselves, but this still constitutes a means of identifying conscious states (via their correlates), which is what the two paradoxical claims relate to.

We might re-evaluate the prospects for property dualism in light of attempts to describe consciousness by reference to fundamental mental properties. This is because, for the property dualist, we must be able to describe our mental lives by reference not to fundamental physical properties but to the set of rules dictating which non-physical properties should be present under which circumstances. This has been done by reference to what Gregg Rosenberg called "the fundamentality thesis," which claims that "our world has another *fundamental* aspect that we must understand if we are to understand the qualitative character

of our mental lives" (Rosenberg 2004: 91). He used this to justify his adoption of *panexperientialism*, the view that experience pervades the whole world. By this view, qualia are present in even the simplest physical entities and when these bear suitable relations to one another, those qualia "merge… into one subject of experience" (ibid.: 92). These properties as existing independently of an experiencing subject are what Rosenberg called "protoconscious" (ibid.: 96), as opposed to Chalmers' "protophenomenal" properties, which are nonphenomenal properties that together form phenomenal properties (Chalmers 1996: 127). Broadly, his position is one of the numerous recent attempts to describe the physical world and consciousness according to Russellian monism.

[6.3] Russellian monism

Russellian monism is a panpsychist or panprotopsychist position premised on an idea argued for by Russell that qualitative properties could be the intrinsic nature of matter (Russell 1927: 345–346). Russell's views on the qualitative nature of matter were brought into contemporary mainstream philosophy of mind when Lockwood defended a position he describes as "Russellian," whereby phenomenal properties are the intrinsic qualities of our brains (Lockwood 1989: 160), before Chalmers went on to state that Russell viewed phenomenal properties as being the intrinsic nature of matter (Chalmers 1996: 154).

In Russellian monism, the intrinsic nature of matter is distinguished from its extrinsic nature, which is that responsible for its causal-structural nature. Russell stated that we cannot understand the physical world purely in terms of properties that bear certain causal-structural relations to each other in a certain way as this explanation would always pass the buck without telling us what any of the related items are:

> There are many possible ways of turning some things hitherto regarded as "real" into mere laws concerning other things. Obviously there must be a limit to this process, or else all the things in the world will merely be each other's washing. (ibid.: 325)

This means that the content of causal-structural statements cannot be determined unless we have some other means of fixing that content. The qualitative intrinsic nature of matter is posited as just such a possibility.

The need for such an explanatory framework is not universally agreed upon. Rorty argued that we should not try to distinguish between the intrinsic and extrinsic properties of something (Rorty 1999: 50) on the grounds that attempting to understand reality as somehow lying behind its appearance is hopeless (ibid.: 49). His antiessentialist alternative is that we think of things such as "tables, stars, electrons, human beings, academic disciplines, social institutions, or anything else" in the same way we think about numbers (ibid.: 53). With numbers, Rorty stated, we have no understanding of their intrinsic nature apart from our understanding of their relation to other numbers (ibid.: 52–53). With regards to objects, sentences regarding appearances, such as "X looks yellow," simply have a distinct usefulness to sentences regarding the way something "really" is, such as "X is blue." It is not the case that one sentence portrays the fact of the matter and that the other portrays only how things seem, it is just the case that describing something as being "really" otherwise than as it appears is simply to state that it is more useful in certain circumstances to describe it as being, for instance, blue instead of yellow. (ibid.: 51).

Rorty's denial of the distinction between intrinsic and extrinsic properties, along with the denial of the distinction between appearances and reality, formed part of his pragmatist project to describe all descriptions of nature in terms of their social function (ibid.: 48). This coupled with his denial of the representational function of language (ibid.: 50) formed the thesis that we never reach truth through forming better metaphysical descriptions. This ostensibly

does not involve any positive metaphysical picture and so there is no need here to show how Rorty's position fares in relation to the Consciousness Science Paradox, but it is worth mentioning just to show that there are means by virtue of which philosophers have denied the intrinsic-extrinsic distinction taken for granted by Russellian monists.

Before discussing the problems that arise once we do adopt such a position, it should be noted that Russell did not maintain any of the prevailing versions of Russellian monism, which is clear from his rejection of the idea that subjects are metaphysically substantive entities. Russell argued that the subject "appears to be a logical fiction, like mathematical points and instants… introduced, not because observation reveals it, but because it is linguistically convenient and apparently demanded by grammar" (Russell 1921: 141). He further stated that there is no good ground for assuming that subjects exist (ibid.: 141). He also expressed the viewpoint that experience is not non-existent, but cannot be divided into consciousness and content, and thus that consciousness is not real (ibid.: 25), which is a viewpoint that he traced back to James (James 1904a: 480). On the contrary, panpsychist variants of Russellian monism assume that subjects pervade the whole of nature, and non-panpsychist variants such as that argued for by Pereboom assume that we must posit as the intrinsic nature of matter that which can account for our phenomenal consciousness (Pereboom 2015: 319). Neither perspective is consistent with Russell's viewpoint.

Goff and Chalmers have categorised Russell as a "panqualityist" (Goff 2017: 160; Chalmers 2015: 271), a view that they both also attribute to Coleman (2012). I will discuss Russell's position in more detail in the following section. For now, it will suffice to state that the absence of the neutral element of Russell's monism is what makes it confusing to categorise Russell's position together with the rest of Russellian monism. Russell stated both that "matter is not so material and mind not so mental as is generally supposed" (Russell 1921: 36); his position sought to buy the advantages of both an idealist perspective and a materialist perspective by conceding elements of widespread understanding of both mind and matter, and for these reasons he defended a neutral monist position (Russell 1927: 382–393). We

misunderstand the nature of both mind and matter by supposing that either is the true substance from which the other is composed, when both are actually constructions that we derive out of a neutral substance from which both are constituted (Russell 1921: 307).

Russellian monism as defended by its contemporary advocates seeks no such middle ground. The concessions it seeks are only to be made regarding our understanding of matter in order to make it more compatible with intuitive insights about consciousness. In his 2017 book defending Russellian monism, Goff argued that the picture of matter painted by Galileo and thus the understanding of the world described through physics is too austere to allow us an understanding of consciousness (Goff 2017: 135–137). While granting a healthy scepticism of accepted perspectives on the nature of matter, he began the book by denying the same thing regarding consciousness and introducing the following constraint:

> *The Consciousness Constraint* – Any adequate theory of reality must entail that at least some phenomenal concepts are satisfied. (A concept is satisfied when it truly corresponds to reality, for example, the concept of God is satisfied if and only if God exists.) (ibid.: 3)

Our intuitions about the existence of consciousness are thus not open to reinterpretation. Given that this one-sided approach entails that Russell's perspective on mind, his subsequent perspective on the intrinsic nature of all matter, and his perspective that neutral entities compose the whole of reality are all abandoned it seems rather misleading to refer to these contemporary perspectives as "Russellian" at all, rather than simply utilising a small set of Russell's arguments to develop a separate set of positions. I will nonetheless follow recent convention in referring to these perspectives as Russellian monist and will classify Russell instead as a neutral monist.

There are three major problems with Russellian monism that will need outlined. The first follows from the premise that intrinsic properties are distinct from causal-structural

properties of matter. If this is so, then Russellian monism suffers the same problem as epiphenomenalism; the fact that we can make accurate claims about the qualitative properties we are aware of must mean that there is some causal connection between those properties and our capacity to make claims. The trouble is that intrinsic, qualitative properties are defined in contrast to causal-structural properties of matter, and so there is no mechanism by which the former can impact the latter.

Howell expressed the similar exclusion argument against Russellian monism (RM) as follows:

> Where an RM property is a property that has a phenomenal categorical ground and some causal dispositions:
>
> 1. there are two distinct and separable aspects of RM properties, those that ground phenomenal resemblance relations and those that ground resemblance between causal profiles;
> 2. all physical events have sufficient causes in virtue of those aspects that ground resemblances between the causal profiles of RM properties.
>
> Therefore, the aspects of RM properties that ground phenomenal resemblances make no unique causal contribution to the physical world. (Howell 2015: 32)

Alter and Coleman have argued that Russellian monists have at least three responses to this argument (Alter & Coleman: forthcoming). Firstly, they can deny that intrinsic properties are distinct from causal properties; instead there can simply be one property and a law governing how that property behaves. Secondly, they can argue that causal properties would not even exist were it not for their intrinsic properties. Thirdly, they can argue that having a unique causal power is not a necessary condition of being causally efficacious (e.g. as argued

by Wilson (2011), something with a distinctive collection of powers could be argued to be causally efficacious (ibid.: 135)).

While these points are arguably sufficient for intrinsic properties to be causally efficacious in some sense, no mere causal efficacy will do. For me to make true claims about the intrinsic properties of something as opposed to true claims about its extrinsic properties, there must be some identifiable aspect of the former specifically as opposed to the latter. They must be able to produce the causal effect of me stating truthfully that I perceive greenness as opposed to redness, and yet if greenness can only causally impact my capacity to make claims by virtue of, for instance, the spatio-temporal relations between a vast number of entities with certain masses and electrical charges, then this offers no plausible mechanism for how I can conceptually distinguish the greenness from that neural stimulation. It is certainly conceivable that the same neural collection could have a different intrinsic nature and yet have the same effect. While the Russellian monist would undoubtedly wish to argue that the same neural collection could not have a different intrinsic nature, the point is simply that the intrinsic nature of matter does not produce any causal effect that can be distinguished from its extrinsic, relational properties, and thus there is no mechanism by virtue of which we would be able to produce a reliable indicator of the presence of these intrinsic properties. As such, if we can make accurate claims about these intrinsic properties, we must thus be doing so purely by chance. This would make all communicable facts about consciousness, including that Russellian monism is true, only possible to make by coincidence.

The second problem of Russellian monism is that it makes the fact that science can teach us anything at all inexplicable. The central premise is that the physical sciences only tell us the relational properties of matter, but that since these relations are only defined by reference to other relations, these references are empty. As such, the facts that we seem to know through science about the constituents of the physical world are simply relational facts with no *relata*. If this is so, then it seems difficult to make sense of our ability to utilise these facts for our greater predictive power and greater precision in manipulation with relation to the

physical world. We cannot be doing this by virtue of the intrinsic properties of that matter given that the only intrinsic properties we are aware of are those belonging to the matter that we are composed of. As such, I am unable to refer to chemicals in a laboratory based on their extrinsic, relational properties, given that these require reference to other relational properties *ad infinitum* and I am equally unable to refer to those chemicals based on their intrinsic properties since I cannot know what these are. We could seemingly avoid the problem by denying that there is a distinction between the intrinsic and extrinsic properties of matter, but this option is not available to the Russellian monist whose position depends upon the intrinsic properties of matter having certain attributes (i.e. their qualitative nature) that do not apply to the extrinsic properties (i.e. to their causal-structural nature). My knowledge of the physical world thus seems inexplicable given Russellian monism.

The third problem is that Russellian monism is clearly victim to the Consciousness Science Paradox. If there are intrinsic qualities of matter then we must be able to determine which qualities relate to which causal-structural properties such that we can state that a certain neurophysiological state has a certain set of intrinsic qualities and causal properties, yet this would require us to be able to empirically identify such relations, which we have established is impossible.

Russellian monism is thus not a viable metaphysical perspective on the nature of consciousness nor the physical world.

[6.4] Summary

I have attempted to show how the Consciousness Science Paradox is an issue for a variety of widely held philosophical positions. Physicalists, property dualists and Russellian monists alike must acknowledge the difficulty if they hope to develop positions that are not susceptible to paradoxical results. We saw that whether we accept a reductive or nonreductive

view, we still must maintain that certain instances of conscious states are identical to certain instances of physical states, in which case we should expect these states to be describable in terms of their causal efficacy, and thus we should be able to empirically determine whether consciousness is present or not. Nonetheless, we cannot [6.1]. We saw that placing consciousness outside the realm of the physical did nothing to alleviate these concerns since, to avoid lapsing into the epiphenomenalism we rejected in [1.2], we are still required to posit that conscious states are causally efficacious and thus the same problem arises [6.2]. We saw that Russellian monists do not avoid the problem by arguing that consciousness is the intrinsic nature of the physical because this involves the denial that this intrinsic nature has any specifiable causal effects, meaning that we once again end up with the same difficulties in accounting for our capacity to produce behaviour in response to conscious states as we do with epiphenomenalism [6.3].

We may seek to evade the paradox entirely by denying the existence of consciousness altogether. As already mentioned, Russell and James argued against the existence of a substantive metaphysical entity such as consciousness, although this denial is part of a very different metaphysics from that which follows from the eliminative materialists, as we shall explore in the chapter 7.

To begin with, it will be worth showing why the panqualityism of recent philosophers such as Coleman cannot be equated with neutral monism to be able to better characterise the position of the neutral monists.

7

Denying the Existence of Consciousness

[7.1] Panqualityism versus neutral monism

Mach argued that the metaphysical distinction between appearance and reality is exaggerated in philosophy (Mach 1897: 11). He focused his attack on the supposed difference between that which appears to the ego and that which constitutes the external world, arguing that the same thing can be viewed as a physical object or psychological object depending on the circumstances:

> A color is a physical object as soon as we consider its dependence, for instance, upon its luminous source, upon other colors, upon temperatures, upon spaces, and so forth. When we consider, however, its dependence upon the retina, it is a psychological object, a sensation. Not the subject matter, but the direction of our investigation, is different in the two domains. (ibid.: 17–18)

Instead, Mach stated that the "the ego must be given up" (ibid.: 24). He dismissed the notion that the ego is a "real unity" and instead regarded it as "a practical unity" or "a more strongly cohering group of elements" (ibid.: 28) and in this sense there is a clear thread running from Mach to James to Russell. With respect to the self, they all endorsed Hume's claim that there is nothing to ourselves other than collections of perceptions (Hume 1739: 1.4.6), which is suspected to have origins in Buddhist thought (Gopnik 2009).

Their viewpoints all share the common elements that a defining feature of the mind, be it consciousness, subjecthood or the ego, are undermined in their metaphysical importance,

with the result being that what are ordinarily taken to be contents of the mind, such as sensations and sense-data, are regarded as constituents of the world at large. Contrasting conceptions of mind and matter are abandoned in favour of a conception of common entities that compose both mind and matter.

That this can only be done by undermining the metaphysical status of the mind can be shown by comparing the traditional neutral monism of Mach, James and Russell to the recent panqualityism of Coleman.

Coleman's perspective is that phenomenal qualities can exist outside the confines of a conscious subject. He adapted an argument used by Foster regarding sense-data (Foster 2000) and argued that it is logically incoherent for something's existence to be exhausted by episodes of awareness by subjects. His argument is that this would entail one of two things. The first is that such a thing derives its existence from figuring in such awareness, in which case "the fact of the existence of the item is constituted by the fact of its figuring in the awareness of a given subject of experience" (Coleman 2012: 151). This would mean that the thing derives its existence from a fact that concerns itself, which Coleman argued, since an item's existence is logically prior to its figuring in awareness, is incoherent. The second is that such a thing's existence takes the form of figuring in an episode awareness, which violates our conception of awareness since awareness involves a subject and object. "A subject could not be said to *become aware of* an instance of phenomenal redness if the phenomenal redness did not exist except in so far as the subject was already aware of it" (ibid.: 151). Both options, then, were rejected by Coleman as incoherent.

Coleman appears to have been following in the footsteps of the neutral monists by denying assumed features of the relationship between experienced qualities and subjects, but his position diverges from these traditional perspectives when, rather than denying that there are metaphysically substantive entities such as subjects, he attempted to use this framework to build an explanation for the existence of subjects.

He argued first for the possibility of phenomenal blending, mixing, and overlapping given relative qualities and positions of entities instantiating such properties (ibid.: 157). He argued that we are familiar with such combination in drinking a decent red wine with Sunday roast beef because the two flavours "pleasingly interpenetrate" (ibid.: 158). He then expanded on this picture by positing that phenomenal qualities can form a "phenomenal multitude" that is responsive to states internal and external to it, "through being able to represent, through being structured" (ibid.: 158). Such a structure should allow for the quality of the phenomenal qualities constituting it to be affected through modification via the senses and cognitive interactions and thus we have a subject.

Subjects, Coleman hypothesised, are able to represent phenomenal qualities outside themselves "by the first item taking on the phenomenal quality of the second" (ibid.: 159). As such, the reason we can have a picture of the reality lying outside of ourselves is that we create an inner copy of the phenomenal array of the world.

By this point, Coleman has diverged significantly from neutral monism. The difference between asserting the existence of subjects and denying it is important not only for the neutral monist picture of ourselves but also its depiction of the nature of reality.

This can be shown by understanding what neutral monists take to be the constituents of reality. Mach stated that "thing, body, matter, are nothing apart from the combination of the elements,–the colors, sounds, and so forth–nothing apart from their so-called attributes" (Mach 1897: 6–7) and stated that space and time may be appropriately labelled as "sensations" (ibid.: 8). James stated that "there is only one primal stuff or material in the world, a stuff of which everything is composed," which he called "pure experience" (James 1904a: 478). Russell, when outlining his own neutral monist position, stated that physical objects are inferred from perceptual events (Russell 1927: 247) and that we could best make sense of these events as having the same sort of qualities as appear to us in perception (ibid.: 345–346). He stated that of the events closest to us in the causal chain leading to our perception, "we know nothing

except what follows from the fact our percepts and “mental states” are among the events which constitute the matter of our brains” (ibid.: 322).

Given, then, that all three, as well as Coleman, agreed that these things are not dependent for their existence upon a subject, they must have had a story to tell about how these things can exist without such dependence. Coleman’s goal was simply to provide us with an understanding of physical entities that would allow them to give rise to consciousness. Specifically, it was to solve the combination problem (Coleman 2012: 138). As such, the independent existence Coleman attributed to the qualities of matter is the independent existence any physical object has under a standard physicalist perspective; the objects exist whether or not they are perceived. To make sense of how subjects formed wholly out of phenomenal properties, as Coleman described them, can know about the world, he hypothesised that phenomenal properties can replicate the quality of another and thus come to represent it (ibid.: 159). As such, Coleman retained the distinction between what lies inside the subject and what lies outside the subject, even if this distinction is numerical and not qualitative.

The case is different for the neutral monist, who attempts to tell a story about how we can come to know about a world composed entirely out of neutral entities. For Mach and James, for instance, there is no distinction between the properties of our experience and those of the world; by this I mean that the two are not only qualitatively identical but are also numerically identical. As already quoted, Mach stated that colours can be considered physical or mental depending on whether we focus on its dependence upon our physiology or its dependence upon surrounding spaces, temperatures and so forth (Mach 1897: 17–18). James stated that the distinction between the “knower” and the “known” is often simply “the self-same piece of experience taken twice over in different contexts” (James 1904b: 538). This means that the subject-object relationship must be replaced by another relationship that can take place between sensations or experiences. James gave the example of himself sitting in the library at Harvard and thinking of the hall. He wrote:

> If I can lead you to the hall, and tell you of its history and present uses; if in its presence I now feel my idea, however bad it may have been, to be *continued*… why then my soul was prophetic, and my idea must be, and by common consent, would be, called cognizant of reality. That percept was what was *meant*, for into it my idea has passed by conjunctive experiences of sameness and fulfilled intention. Nowhere is there jar, but every later moment matches and corroborates an earlier. (ibid.: 539)

James' point, then, was that there is no need for there to be metaphysical entities that fill either the role of subject or object; rather we can simply have experiences or sensations that can, by bearing certain relations to one another, *mean* or *know* one another.

Russell's perspective seems on the face of it to be distinct from that of James and Mach, but in certain respects, particularly with regards to the *neutral* element of neutral monism, it has much more in common with the latter. While Russell followed James and Mach in his claim that there is no distinction between a noise and the hearing of a noise (Russell 1918: 255), his commitments to the causal theory of perception do not seem to have permitted him to fully endorse an identity between the qualities of a percept and those of the object that cause the percept. He stated, "The space of physics is connected with causation in a manner which compels us to hold that our percepts are in our brains, if we accept the causal theory of perception, as I think we are bound to do" (Russell 1927: 383). However, it is his denial of the existence of the subject that formed the basis for the view that we derive both the existence of minds and that of matter from the presence of perceptual qualities. He considered the case of thinking that the existence of a patch of colour of a certain shape and our seeing it are two separate things:

> This view, however, demands the admission of a subject… If there is a subject, it can have a relation to the patch of colour, namely, the sort of relation which we might call awareness. In that case the sensation, as a mental event, will consist of awareness of the colour, while the colour itself will remain wholly physical and may be called the sense-datum, to distinguish it from the sensation. The subject, however, appears to be a logical fiction, like mathematical points and instants… we must dispense with the subject as one of the actual ingredients of the world. But when we do this, the possibility of distinguishing the sensations from the sense-datum vanishes; at least I see no way of preserving the distinction. Accordingly the sensation that we have when we see a patch of colour simply is that patch of colour, an actual constituent of the physical world, and part of what physics is concerned with. (Russell 1921: 141–142)

The parallel between Russell's view and Mach's view is clear. How I understand Russell's view is as follows. Physics suffers a major problem if it cannot account for perception, given that it is only via perception that physics is possible (Russell 1927: 137). What we perceive is data, and it is on the basis of this that perceptual inference is possible (ibid.: 187–192). We only infer the existence of external objects from this data (ibid.: 197) and there is no reason for us to infer from empirical observation the existence of a subject of experience (Russell 1921: 17–18). From the data, we have good reason to suppose that external objects exist, which is that positing them enables us to make sense of regularities in that data as well as predict somewhat accurately which data will occur together (Russell 1927: 198–199). From correlating the data about the external world with data about other people, such as their reports to be perceiving certain things, we derive that there are objects that produce common perceptions in multiple people (ibid.: 207). This gives us a worldview featuring physical objects that induce perceptions in us via causal interaction with us, but ultimately the entire world is

inferred wholly from sense-data. This inferred world consists of entities with certain relations to each other, but we do not infer anything about their intrinsic nature (ibid.: 325–326). Yet, the fact that these relations exist suggests that there must be items that can be described in terms of their actual, non-relational nature (ibid.: 325), and there is no better candidate than the only data we have about reality; the same sort of sense-data by virtue of which we derive our entire understanding (ibid.: 346).

When Russell stated, then, that a patch of colour is an actual constituent of the physical world (Russell 1921: 142), his commitment to the causal theory of perception required him to accept that it exists in the head of the perceiver, not as a mere appearance, but as a physical event partially forming the matter of the perceiver's brain. It is only from the presence of such percepts that a person can then attribute structural properties about the rest of reality, and then it is our knowledge of the intrinsic properties of our brains that allows us to infer that the rest of the world may have such intrinsic properties.

It is in such claims regarding the existence of the world outside of all possible perception that Russell went beyond James, whose radical empiricism gives him no cause to go beyond our experience. For James, his description of how two people can know the same fact involves a single unit of experience existing in the perception of two people simultaneously (James 1905: 178) and so he avoided the need to infer the existence of physical objects except insofar as they figure in the experience of people.

I hope it is clear now why Coleman's view cannot be said to be the same as Russell's. Russell, like James and Mach, regarded the data of our sensations as the sole constituent of all reality and the fundamental basis of all our knowledge *including* our knowledge both of ourselves as subjects and of physical objects. Coleman, on the other hand, took the existence of physical objects and subjects of experience as the data that require explanation; the qualities of the world are posited in order to make it possible for subjects and objects to inhabit the same world. Neutral monists' theories are only designed to make sense of how we can build

knowledge out of the qualities we experience, even if this means sacrificing the previously accepted reality of both subjects and objects.

The difference between Coleman's perspective and that of the neutral monists is important to draw here because it can be demonstrated that Coleman's perspective is straightforwardly subject to the Consciousness Science Paradox in a way that the neutral monist's position is not. For Coleman, the constituents of the physical world have phenomenal properties that combine in certain ways to form subjects of experience. The result of this is that a physiological system such as a human body can, supposedly, be imbued with phenomenal properties that can coherently constitute our consciousness. Of course, if this is the case then there are two contradictory facts that must follow. One is that the relationship between our physiological states and the phenomenal properties that inhabit them must be possible to empirically determine. The other is that this relationship is impossible to empirically determine. As such, Coleman's panqualityism is victim to the same problem as the other Russellian monists.

The problem with neutral monism will require further explanation.

[7.2] The problem with neutral monism

Russell's neutral monism runs into trouble primarily not because of his attempt to build a model of the world out of qualitative properties but because of his commitment to the world that he believes we must infer from those properties. I have already outlined in chapter 4 why the causal theory of perception results in a situation where we have to assume that our perceptions are simultaneous with or else immediately preceded by a physical stimulus and this assumption is not amenable to testing, but it will pay to show, having outlined Russell's perspective in more detail, why it causes a problem for his particular worldview before moving on to a major concern with neutral monism generally.

Russell stated that “what the physiologist sees when he examines a brain is in the physiologist, not the brain he is examining” (Russell 1927: 320). This picture of reality, then, although it does not involve subjects, and although the world it describes is inferred from a neutral substance, does not in any way avoid the main problem of the Consciousness Science Paradox. The world that Russell described is one in which percepts are induced in an individual by a causal chain that ends “in our heads” (ibid.: 320). This means that these percepts are assumed to correlate with particular physiological states. Indeed, the percepts are assumed to constitute our physiological states. Russell may appear to avoid the problem because percepts are the means by which we make empirical observations and inferences about the physical world. Yet, these inferences lead us, according to Russell, to the conclusion that the percepts exist inside our heads. The inferences, then, lead us straight to a problem very similar to the Consciousness Science Paradox; we must be able to empirically correlate the presence of these percepts with the presence of certain physiological activity, and yet there is no possible way to make such identifications.

A demonstration of how the paradox is problematic for James’ view will require further explanation because he did not commit to the causal theory of perception as Russell did. In fact, James accepted that the experience we have in seeing an object can be numerically identical to the experience constituting part of a physical object with only the relation of that experience to other experiences allowing us to count it once as “thought” and once as “thing” (James 1904a: 480). In that sense, what we perceive is not in our heads at all, but rather has a range of relations to various other perceptual qualities, including close spatial relations with the other properties making up physical objects. My perception of a cube shape thus fits into a psychological story about what other thoughts this brings to mind and memories I have of similar-shaped objects, but it also fits into a physical story of being a cube-shaped object with a certain mass that occupies a certain region of space at a certain time.

It is in his conception of the temporal relations between the physical and the psychological that the particular difficulty I wish to discuss in relation to James is produced.

James stated that the relations of “thought and thing” to time “are identical” (ibid.: 488). Although he accepted that percepts considered as psychological features and percepts considered as physical features may be numerically identical, he did not consider the possibility of numerical identity between “a clear image” before his mind of Memorial Hall while he sat in the library and the appearance of Memorial Hall when he actually stood in it (James 1904b: 539). Indeed, he considered them “two pieces of *actual* experience belonging to the same subject, with definite tracts of conjunctive transitional experience between them” (ibid.: 538). The relationship between the two was made by reference to “relations that unroll themselves in time” (ibid.: 539). He stated:

> Whenever certain intermediaries are given, such that, as they develop towards their terminus, there is experience from point to point of one direction followed, and finally of one process fulfilled, the result is that their starting point thereby becomes a knower and their terminus an object meant or known. (ibid.: 539–540)

It is clear, then, that James considered events occurring in our psychology to be temporally located alongside physical events. In this, he followed Mach, who claimed, “That a definite, specific time-sensation exists, appears to me beyond all doubt” (Mach 1897: 248) and that it “accompanies every other sensation, and can be wholly separated from none” (ibid.: 245). He further stated, “For all time-sensations, also, I must suppose like nerve-processes” (ibid.: 62), which, given that nerve-processes are themselves spatiotemporally located, entails that the time-sensations themselves have spatiotemporal correlates.

Russell also agreed, claiming that percepts can be “compresent” with physical events (Russell 1927: 383–384). Indeed, he used temporal simultaneity as a justification to locate our percepts in our heads:

> I hold, therefore, that two simultaneous percepts of one percipient have the relation of compresence out of which spatio-temporal order arises. It is almost irresistible to go a step further, and say that any two simultaneous perceived contents of a mind are compresent, so that all our conscious mental states are in our heads. I see as little reason against this extension as against the view that percepts can be compresent. (ibid.: 385)

Neutral monism thus consists of the view that, while the qualities making up the world can be numerically identical to those occurring in our minds, mental events must correspond with physical events in time.

The difficulty this produces is in characterising what the causal efficacy of these mental events are. You can imagine a room bursting into flames, but this need not involve any room (real or imaginary) actually getting burnt (James 1904a: 482), but to state truthfully that the percept of a flame is present at such a moment, there must be some property present at that time by virtue of which I can make such a claim. Yet this property must be able to have a physical effect, which means that we should be able to empirically identify the physical correlates of this property. This lands us at the foot of a problem equivalent to the Consciousness Science Paradox. It does not matter that our physiological state, for instance, is made of percepts; we still need to be able to identify which of the percepts we have in making empirical observations of physiological states, such as in visually perceiving neurons, correspond to the percepts we have in instances of perception, such as what we see when we look out at a landscape. Yet, doing so is impossible. It is impossible because no empirical observation can tell us which percept we have in empirical observation corresponds to which percept we have by virtue of being in a physiological state. The reasons for this are the same as those for why we cannot correlate physiological states with phenomenal properties; no empirical observation reveals such a correlation.

Neutral monism takes percepts, or experiences, or sensory qualities as both the metaphysical foundation of the world (James 1904a: 478; Russell 1927: 346) and the epistemic foundation from which we build our knowledge of the world and ourselves (James 1904a: 482; Russell 1927: 187–192). It is because neutral monists have denied that there is a numerical distinction between the qualities lying within our awareness and the entities that are in a state of awareness that they have then gone on to argue that consciousness does not exist. For James and Russell, consciousness is simply a theoretical entity used in our descriptions of reality and they have both argued that the inclusion of this entity in these descriptions is a mistake. James argued that the existence of consciousness supposes that experience is "indefeasibly dualistic in structure" (James 1904a: 478), whereas he took it to be actually true that the separation of consciousness and content comes "not by way of subtraction, but by way of addition," whereby epistemic relationships lie between experiences, one of which features as "knower" and one of which features as "known" (ibid.: 480). Russell argued that the existence of the subject of experience supposes that there is an entity that can bear the relationship of "awareness" to a percept, whereas in actual fact that there are simply percepts and it is from the relationships of various percepts that we construct the existence of the "logical fiction" of a subject (Russell 1921: 141–142). As such, it is in the described epistemic primacy of the experiential quality and in opposition to what is seen as an imagined structural complexity within experience that neutral monism can be considered a perspective that readily denies the existence of consciousness. Nonetheless, for neutral monists there remains a picture of mentality just as liable to fall victim to the same sort of paradox.

The problem with neutral monism arises as a result of the fact that, in denoting that the qualities that constitute the world are data, its basic elements of reality are defined by our capacity to perceive them from moment to moment. This should not be surprising for James, given that he used neutral monism as the metaphysical basis for his radical empiricism, which is the view that "experience as a whole wears the form of a process in time, whereby innumerable particular terms lapse and are superseded by others that follows upon them by

transitions which, whether disjunctive or conjunctive in content, are themselves experiences" (James 1904b: 541–542). Nor should it be surprising for Russell, given his commitment to the causal theory of perception. The neutral monist is committed to the perspective that we build our conception of the world from a continuous flow of data, and for this flow to exist at all, the percepts present to us must change over time. Yet, this cannot be reconciled with the fact that it is impossible to empirically observe which physical states correspond to which of these percepts.

[7.3] Eliminative materialism

If we are going to deny the existence of consciousness to avoid the paradox, we might instead attempt to avoid the radical reinterpretation of the nature of the physical posited by the neutral monists and argue that our conception of the physical is more or less accurate while our conception of consciousness is wholly inaccurate.

Eliminative materialism, or eliminativism, was introduced earlier in this thesis when characterising consciousness in the first place and so, rather than repeat that description of the position, I will begin by discussing Feyerabend's eliminativism before describing how adopting an eliminativist worldview does not allow us to neatly avoid the Consciousness Science Paradox.

Hints of Feyerabend's eliminative materialism can be found in his discussion of materialism. He opposed the notion that the existence of mental concepts could be taken as an argument against materialism by arguing that statements concerning mental processes lack content, and it is only by virtue of this lack of content that we can take ourselves to be certain about such statements (Feyerabend 1963: 56). His defence of materialism bears a strong resemblance to his anarchism in a certain regard, which is that his main purpose was to attack widely-held philosophical worldviews rather than to systematically build an alternative position.

As Feyerabend argued, it is our language that has developed such that our description of thoughts is incompatible with our description of brain processes. Given that this language developed over "considerable time… the materialist philosopher must be given *at least* as much time" to develop an alternative language (ibid.: 54).

It is at this point not clear whether Feyerabend is defending reductive physicalism or eliminativism, although the distinction between these two is not clear cut in terms of its advocates. As mentioned previously, Smart and Place, two paradigmatic examples of reductive physicalists, rejected phenomenal qualities and so they would have fit comfortably into the ranks of eliminativists. Smart's metaphysics of the mind required only that the concepts we use to denote sensations are simply those that occur under particular circumstances, not those that occur alongside the presence of a particular subjective class of properties. This is similar to the view of Feyerabend, who stated that what is important in a materialistic description of pain is that it describes something with particular causal antecedents (Feyerabend 1963: 58). To describe what pains are, he wrote, we need only recognise information such as that, "they do not reside in tables and chairs; they can be eliminated by taking drugs; they concern only a single human being; they are not contagious" (ibid.: 64). The dispute, then, was not that there is no such thing as pain, but rather that pain can be described by reference to a particular thing without any reference to properties that are "immediately given," to which any reference would be meaningless (ibid.: 64).

Philosophical suspicion of immediately given knowledge formed the main thesis of Sellars' essay titled "Empiricism and the Philosophy of Mind" (Sellars 1956). Sellars argued that to state that something appears a certain way (e.g. looks green to you) is more logically complex than to state that something is a certain way (e.g. is green) (ibid.: 144–145) and so to state that we first learn how to identify how things appear to us and then learn to make claims about the way they actually are is an incorrect description of how we learn to apply concepts of our experiences. Sellars stated that the problem with the thesis that there are immediately apprehensible properties, which he calls "the Myth of the Given" (ibid.: 140), is that to state

that something seems a certain way, such as that you have the “impression of a red triangle”, is to state that you have an “impression of *the sort which* is common to those experiences in which we either see that something is red and triangular, or something merely looks red and triangular, or there merely looks to be a red and triangular object over there” (ibid.: 175). He concluded from this that such impressions have only an “ostensive definition”, meaning that the concepts we apply to the contents of our own experiences are purely formal and do not denote immediately apprehensible, given properties but rather simply denote that the observer is in the sort of state they would usually be in in the presence of a red and triangular object (ibid.: 193).

Sellars’ concerns echoed those of Wittgenstein, who, as already mentioned in the earlier discussion of verificationism, rejected philosophical descriptions of private entities as genuine parts of language (Wittgenstein 1953: 293). The fact that Feyerabend’s criticism of immediately apprehensible sensory qualities was wholly consistent with such positions is further evidence of Wittgenstein’s influence over his philosophy.

Eliminative materialism is a substantive description of reality that has grown naturally from well-developed philosophical projects that require both proper defence and proper refutation. It makes clear claims about the nature of reality that cannot be dismissed by philosophers who have refused to take it seriously by responding to those asking what qualitative properties of consciousness are with, “If you got to ask, you ain’t never gonna get to know” (Block 1978: 281), or by stating that denying the existence of experience is “the deepest woo-woo of the human mind” (Strawson 2006: 5). Although the reality of phenomenal properties may entail that we simply know from introspection that such properties exist, this does not mean that the only possible situation in which a group of organisms would come to behave as though they have phenomenal properties is that they actually do. It is not enough to dismiss positions that deny phenomenal consciousness only to replace them with other views that are susceptible to the Consciousness Science Paradox and thus cannot be

assumed to be correct, as we found was the case in views adopting the phenomenal concepts strategy for example [6.1.1].

Eliminativism may appear to be exactly the sort of project we should be looking for here as indeed it does give us a means of avoiding the Consciousness Science Paradox, and the reason it fails to provide an alternative to the flawed worldviews that result in the paradox will take some explanation.

Eliminativism consists broadly of the belief that folk psychology is a mistaken theory and that it is perfectly reasonable to assume that we will have no need for it once we replace its terms with neurophysiological concepts. Paul Churchland's argument against propositional attitudes was that they are based on an outdated empirical theory (Churchland 1981). Given such claims, the eliminativist must replace talk of whichever mental descriptors they wish to abandon with something. When somebody screams out in pain, the eliminativist cannot deny that something is happening, and indeed they have no intention of denying this, but they can deny that that person's screams, gestures, and descriptions are indicative of the presence of mental states.

Rather, eliminativists seem invariably to have supposed that all these things are only indicative of the presence of certain states best described by neuroscientists, such as Rorty's suggestion that mental state descriptions can be superseded by descriptions of brain processes if the latter has greater descriptive power (Rorty 1965), although the effort required to change our language such that "sensation" terms are eliminated makes such a transition impractical (ibid.: 32). Nevertheless, he stated that the idea that we could replace talk of sensations with talk of brain processes is "sensible and unconfused" (ibid.: 52). Paul Churchland claimed that what the elimination of folk psychology will involve "depends heavily on what neuroscience might discover, and on our determination to capitalize on it" (Churchland 1981: 84). Feyerabend, in arguing for materialism, argued that there was "not a single reason why the attempt to give a purely physiological account of human beings should be abandoned" (Feyerabend 1963: 65).

Interestingly, in his later *Philosophy and the Mirror of Nature*, Rorty envisioned the mind-body problem being absent for an imaginary race of beings called “Antipodeans” who have learned neurophysiological concepts in place of mental state concepts (Rorty 1979: 70–77). The argument defended an account of philosophical problems as arising as a matter of contingent historical circumstances and so it seems as though the point would have been used to defend Rorty’s earlier eliminativism. He never made such a point though as, like with Feyerabend’s anarchism, the arguments in that book are purely destructive. Instead, as Tartaglia stated:

> We remember, in particular, that Rorty’s motivation is metaphilosophical doubt rather than philosophical perplexity, and that he is somebody who refuses to ‘bow down’ to an ‘authority called Reality’ (Rorty 2000: 376). What actually happens, then, is that he rejects eliminative materialism, advises that the best attitude to adopt towards the mind-body problem is boredom, and ultimately refuses to endorse any positive position on the mind whatsoever. (Tartaglia 2007: 73)

Rorty’s abandonment of the eliminativist project notwithstanding, what would be required for such an accomplishment is for mental state concepts to be fully replaced by neurophysiological concepts as they were in Rorty’s Antipodeans example. The eliminativist’s worldview thus depends upon how we determine whether something is a mental state concept. Eliminativists have sometimes focused on particular mental terms they would wish to see removed, such as “belief” (Churchland 1981), but even in such cases the point where we demarcate the mental state concept from one which we require in our new worldview is not clear.

This is evident from much of the dispute about whether eliminativism is self-refuting (see Reppert 1992). One accusation levelled at the eliminativist was that if they claim that belief does not actually exist, they cannot be said to believe anything, and thus the whole

edifice of epistemology comes crashing down, leaving the eliminativist incapable of justifying anything at all, never mind their own position (Swinburne 1980). If there is no such thing as truth, we might argue, eliminativism cannot be true. This, though, supposes that there is no option to scrap current and historical epistemological models and replace them with entirely new ones and thus the accusation begs the question against the eliminativist, presupposing that their project is doomed to failure (Ramsay 1990).

The option is open to the eliminativist to claim that asserting a lack of viability in a position that denies its own truth begs the question in favour of mental state concepts. Paul Churchland gave an analogous example of a vitalist claiming that an antivitalist's argument against the existence of vital spirit as the basis for life is question begging "for if the claim is true, then the speaker does not have vital spirit and must be *dead*," in which case "his statement is a meaningless string of noises, devoid of reason and truth" (Churchland 1992: 22). We could argue that once we replace mental state terms with neurophysiological ones, perhaps "eliminativism is true" will be replaced by a more accurate and descriptively useful phrase in the new neurophysiological language. The problem is that, while I cannot reject this possibility before such a language has had time to develop, it seems very problematic for the eliminativist to rely upon such a thing. After all, if we can simply have free speculation on future conceptual developments, there is nothing stopping us from supposing that some future conceptual developments will give us a completely non-problematic description of the communication between non-physical minds and brains. Ramsay outlined the state of both sides in this dispute:

> The burden of proof is on the eliminativist to put forth and defend a plausible theory of cognition that does not invoke the posits of common-sense psychology but can none the less do much of the work presently done by folk psychology. And with one or two possible exceptions, they have not yet made good on this claim. But the burden of proof is on the self-refutationist to show

> that these efforts are futile – that developing and defending such a theory is, in some sense, conceptually impossible. (Ramsay 1990: 463)

It is both true that eliminativism cannot be ignored on the mere assumption that folk psychology cannot be superseded and that eliminativists can only counter this accusation by presenting an alternative conceptual picture. Yet, there is good reason to suppose that developing such a theory is conceptually impossible. After all, we have already seen how neuroscience is in principle unable to draw up any correlations between conscious states and physical states, so it does not seem as though neuroscience was a viable candidate for a description of our mental lives even prior to our eliminativist project. For instance, Patricia Churchland argued that customary usage of mental terms such as "remember" are not important in determining, for instance, whether brains can remember, but rather "whether, given the empirical facts, it is a reasonable hypothesis that brains remember" (Churchland 1989: 273–274), with the clear and unjustified assumption that the truth or falsehood of such a hypothesis is possible to determine through empirical investigation. Even if it had seemed this way, eliminativism cannot be justified by correlations between neurophysiology and conscious states since the latter are posited to be non-existent; alleged relationships between the two should not even be factored into the eliminativist's worldview.

It is worth noting here that I am assuming that eliminativists would wish to see conscious states eliminated rather than simply identified with neurophysiological states, although I do not believe that this is clearly the case given that eliminativists such as Patricia Churchland have claimed that the concept of consciousness can be conceived as being a neurophysiological concept (Churchland 1983: 92–93). Nonetheless, for those eliminativists about mental states who believe that conscious states are neurophysiological states, their position is straightforwardly embroiled in the Consciousness Science Paradox, as it relies upon empirically identified correlations between conscious states and neurophysiological states that

cannot be found. So, the only position it is of value to discuss at this point is that which assumes that consciousness does not exist.

For those eliminativists who believe that consciousness does not exist, there are serious problems that can be traced to their reliance upon the adoption of neurophysiological concepts as their proposed replacement for such a concept. Let us assume that a neurophysiological language has been developed such that for every mental state expression there is a superior neurophysiological expression and that the former have been entirely eliminated in favour of the latter (although of course this relationship need not be one-to-one, where each mental state expression is eliminated in favour of a single neurophysiological expression). In such a language, it must be possible to refer to things other than brain states; the expressions in this language must be able to say more than that I am in a particular brain state at a particular time. In Rorty's example, the Antipodeans are supposed to talk both about their brain states and states of the wider world:

> Sometimes they would say things like "It looked like an elephant, but then it struck me that elephants don't occur on this continent, so I realized that it must be a mastodon." But they would also sometimes say, in just the same circumstances, things like "I had G-142 together with F-11, but then I had S-147, so I realized that it must be a mastadon." (Rorty 1979: 70–71)

It is easy to see why references to both are required. Without being able to refer to empirical observation of anything lying outside of my brain, for instance, we would have no way of supporting any of our empirical knowledge. Without empirical knowledge (or whatever we would call it in the neurophysiological language), we have no way of supporting our understanding of the world at all, including our understanding of neurophysiology. We would lose all reference to anything other than brain states, which we could no longer expand our knowledge of nor justify our understanding that such brain states exist.

If, on the other hand, the neurophysiological language would be able to describe not just the brain states we are in but that which we are aware of by virtue of being in those brain states, this would allow us to refer indirectly to the world via our neurophysiological references; to state that I am in brain state G49 would tell you not only my neurophysiological condition but also that I am aware of the sound of crunching leaves. However, if such an expressive language is possible then we have no reason to suspect that statements such as, "The non-physical mind communicates with the physical brain," should not be expressible (in some superior way) in the new neurophysiological language. The possibility of neurophysiological translation alone then does nothing to resolve disputes about folk psychology, given that it seems just as feasible that the same concepts and the same problems will be expressible. Moreover, the desired replacement of mental state terms with neurophysiological terms would be equally compatible with the falsehood of eliminativism as it would be with its truth, since we could argue for a different metaphysical position using those neurophysiological terms. Thus, even if such a language could be deemed plausible, this alone would not justify eliminativism.

Given then, that the possible replacement of folk psychological concepts with neurophysiological ones does not justify eliminativism, it is questionable that we should be looking to neuroscience for such conceptual development at all. After all, eliminativists about consciousness are not permitted to assume that conscious states do indeed correlate with neurophysiological states and avoiding this assumption by denying the existence of consciousness is precisely how they could avoid the Consciousness Science Paradox in the first place. Yet, they immediately renounce this advantage by maintaining that it is the explanatory success of neurophysiological concepts that is placed to supersede our conception of consciousness. For either these neurophysiological concepts make no explicit mention of the content of our awareness, in which case they are inadequate to form the basis of our empirical understanding of anything at all, or they do mention the content of our awareness, in which case we can only establish such a role for these concepts by drawing correlations between the content of awareness and neurophysiological concepts, and this is

simply to become embroiled in an almost identical problem to the Consciousness Science Paradox. This is clear if we simply notice that empirically observed correlations between the content of awareness and neurophysiological concepts are only possible to make if we presuppose that we have some independent measure of the content of awareness, and we do not. If we once believed that recognisable conscious behaviour was such a measure, I hope that the analysis of such behaviour as an empirical indicator already conducted in this thesis has been enough to show that it is not [2.4–3.5].

This leaves a more radical avenue for the eliminativist: she can deny the existence of mental content altogether. While she can claim that we have empirical knowledge of certain elements of physical reality, she can deny that this entails us holding content in our minds at all but rather claim that empirical knowledge consists of nothing more than neurophysiological states standing in certain causal relations to things in the world. Using such a tactic, the eliminativist could respond to accusations that empirical knowledge is not possible in the first place without consciousness by stating that this begs the question against eliminativism.

In such a case, we would suppose that observing, for instance, that a red post-box is in front of me could be described something along the lines of, “I am in brain state Y32.” Of course, the only way I could know that I was in brain state Y32 is if I was in the relevant brain state of having perceived that brain state Y32 is currently present. The question that this raises, then, is this: if I have to independently verify which brain state I am in given a particular empirical observation using *another* empirical observation, then why describe the situation as, “I am in brain state Y32” rather than, “I am seeing a red post-box” at all? That is, given that the possibility of empirical observation must be presupposed for such a project to get off the ground in the first place, there does not seem to be a clear benefit to be found in rephrasing every possible observation in neurophysiological terms, since these are simply other observation terms. Furthermore, if consciousness does not exist precisely because conscious states are unlikely to have any correspondence to brain states, then there is no reason we should wish to replace all conscious state expressions with brain state expressions any more

than those who believe in consciousness should wish to replace all observation statements with claims specifically about consciousness. The fact that the eliminativist places perceptual and introspective knowledge within neurophysiological states is irrelevant; this simply entails that a statement about *how* something is known can be told by reference to a neurophysiological state, not that statements about *what* is known at any one time is told by reference to a neurophysiological state.

Indeed, if the eliminativist conflates what we know about the world with how we know about it such that we are not only aware of the world via neurophysiological states but we are only actually aware of neurophysiological states in the first place, this entails an indirect realism about the world lying outside of our brains with us only directly perceiving brain states. If this is so, and we are immediately apprehending our own brain states, then, given that we cannot tell what these brain states are simply in having them, we have a multiplicity of immediately apprehended brain states on our hands that need to be correlated with empirical observations of neurophysiological states. I hope by now to have demonstrated that this cannot be done.

These considerations could make it seem puzzling that eliminativists place so much explanatory work in the hands of future neuroscientists, but the reason for this is clear once we try to imagine what an eliminativist project would look like that assumed this was not a matter for neuroscience. If instead the eliminativist took empirical observation for granted and gave statements such as, “I am seeing a red post-box” the same level of epistemological certainty as “I am in brain state Y32,” then this supposes that the properties we take ourselves to perceive – the red shape in our vision – is an actual constituent of the external world. They would not be able to accept something like the primary/secondary quality distinction because this requires qualities to be able to exist as mental content, which the eliminativist denies, and so they would have to assert that qualitative properties can exist in the world outside of our brains. This would make eliminativism, perhaps the most metaphysically austere position in philosophy of mind, guilty of a bountiful proliferation of qualitative properties. This is an unacceptable consequence; the resulting position would assuredly not be an eliminative one.

Rather, consciousness would be eliminated only to splash its contents across the observable world.

Even now, it would be possible for the eliminativist to accuse me of begging the question on the grounds that I am supposing that the eliminativist's project cannot work without the idea of empirical knowledge. I now counter that the question-begging objection can only go so far. Let us imagine that we wished to defend the hopelessly flawed metaphysical position that descriptions of all physical and mental states can be eliminated in favour of descriptions of the various states and relations of baked beans. All expressions such as, "I believe in God," or, "There is a cat on the mat," would be argued to fail to refer to an actual state of the world, whereas, "a pile of forty beans stands beside a pile of thirty six beans," would succeed. In this case, we may appropriately ask in what sense this new description is more accurate or more useful, at which point the bean theorist may accuse us of begging the question; to ask about "accuracy," "usefulness," or "truth" is to presuppose that bean theory is false. After all, these expressions can be eliminated in favour of bean terms. We may frustratedly ask the bean theorist how we are supposed to figure out which bean theory terms to use in which situation, at which point the bean theorist may irritably respond that we are, again, begging the question. "Using certain terms in certain situations," the bean theorist may argue, is a phrase that only has meaning in non-bean theory language, and so to suppose that it is a weakness that bean theory is unable to account for this is to presuppose that bean theory is false. Any alleged weakness of bean theory can thus be dismissed as begging the question against bean theory.

The reason eliminativism is nowhere near as patently absurd as bean theory is because eliminativism has the trump card of our current and future empirical knowledge regarding neurophysiological states to play. Indeed, this is supposed to provide a model for the introduction of new neurophysiological state expressions that will supersede our mental state expressions. Whatever mental state expression we wish to use, they can argue that it is better eliminated in favour of expressions that can be found in the language of our empirical knowledge of neurophysiological states. However, if they undermine the very idea of empirical

knowledge, denying the existence of belief, truth, justification, and first-person observation, then they are arguing that the very rules of the game are hopelessly flawed, in which case it is anybody's guess as to whether their trump card is valid. If the eliminativist was to argue simply that it is begging the question to state that we need empirical knowledge to refer to neurophysiological states, then it is extremely unclear as to what the benefit of using neurophysiological expressions is. If they are not true, or justified, and we have no empirical knowledge of neurophysiological states, then what is the purpose of adopting expressions that refer to neurophysiological states at all? My challenge is to ask how this same line of reasoning could not be used to justify bean theory.

The better option for the eliminativist, I suspect, is to argue that we have empirical knowledge of a world consisting wholly of non-qualitative properties. The trouble is that the misapprehension that there are qualitative properties cannot be attributed to a distinction between the way the world appears and how it actually is because this requires an appearance of the world by virtue of which we are aware of it, in which case we would have more directly apprehensible properties to correlate with neurophysiological states. The misapprehension must then be attributable to brain states, resulting in the contemporary illusionist position that we simply judge ourselves to be in certain conscious states.

[7.4] Illusionism

The more recent eliminativist project of illusionism suffers from a similar difficulty. Illusionism is an eliminativist project because it denies that qualitative aspects of consciousness exist (Frankish 2016: 21). Psychologist Nicholas Humphrey has argued that phenomenal properties are a "kind of make-believe" (Humphrey 2011: 32), Hall argued that we can explain our representations of phenomenal qualities without needing to posit that there actually are such things (Hall 2007: 209), Rey argued that all we need to explain in our

phenomenal judgements is the functional role such judgements play rather than phenomenal properties themselves (Rey 1992), and Frankish has argued that it is more difficult to account for our judgements about qualitative properties by referencing actual qualitative properties than it would be by simply explaining how we could come to make such judgements in the absence of such properties (Frankish 2016). Dennett stated that he does not deny that consciousness exists but believes that it is different to what people think it is (Dennett 2018).

Illusionism thus seeks the explanatory advantage of avoiding metaphysical commitments to phenomenal properties and any philosophical difficulties we have in our conception of these while still being free to describe how it is that we make judgements about such properties. By describing how we make such judgements in non-phenomenal terms, we can provide an explanation of our beliefs about consciousness without having to answer difficult philosophical questions about what causal effects phenomenal properties have if any (Frankish 2016: 27).

Tartaglia disagreed that this conception of consciousness can make sense of our judgements about experience. He argued that it is possible for us to make systematic false judgements about the presence of a tree, for example, provided we explain this by reference to something like people being hypnotized or there being something causing hallucinations in the area where a tree is experienced as being. He argued that this explanation works because there is a distinction between the experience of there being a tree and the presence of an actual tree such that we could still have the former even in the absence of the latter. Yet, this move cannot be made by the illusionist because they maintain that our phenomenal concepts do not refer to anything that exists, and as such there is nothing that persists when we falsely judge a tree to be present other than our judgement that it does. This entails that it is not by reference to an understandable mistake, as we might be making if we perceive an appearance that unknowingly to us at the time does not match reality, that the illusionist can explain our perceptual illusions, since such appearances do not actually exist, and as such our false judgements must be “baseless and inexplicable” (Tartaglia 2016: 93). Tartaglia’s objection

was thus that illusionism (or as he called it, “revisionism” (ibid.: 89)), maintaining as it does that all phenomenal judgements are similarly baseless, can make no sense of our inclination to judge that we have experiential properties in the first place (ibid.: 94).

The illusionist might retort that phenomenal concepts do not need to be accompanied by the presence of some actual phenomenal appearance to serve the functional role they do. Indeed, the option is open to them to agree that our false judgements are baseless and inexplicable, and to argue that phenomenal consciousness is nothing other than the inclination to make such false judgements. Humphrey argued that phenomenal concepts are a way of certain parts of our brains telling other parts that something is important. This “something” is either an environmental stimulus or something internal, like a bodily state or even our own life. Humphrey gave the example of us taking pleasure in simply being alive:

> We accept that Nature made sex pleasurable so as to encourage animals to take the steps that lead to sexual intercourse. Then why not make the feeling of existence magically delightful in order to encourage conscious creatures to do the things that lead to their existing? (Humphrey 2011: 85)

In response to the question of why natural selection could not have provided us with judgements of this kind without making us believe in phenomenal consciousness, Humphrey stated that this illusion “gives you (or at any rate gives you the illusion of) a *substantial thing* to value” (ibid.: 87). He referred to psychoanalyst Ernst Becker’s statement that humans attempt to avoid the fatality of death “by denying in some way that it is the final destiny of man” (Becker 1973: xvii) and stated that fear of death must thus be “highly visible to natural selection – and hence so must have been the consciousness that lies behind it” (Humphrey 2011: 94). The function of consciousness is thus, according to Humphrey, not to give us accuracy in our judgements but to provide a survival advantage.

The illusionist thus removes the phenomenal picture of consciousness where consciousness is characterised as being a locus of phenomenal states, and she replaces it with one of two pictures. Either she denies the existence of consciousness, declaring it illusory (Rey 2016: 197), or else she argues that consciousness does exist but that phenomenal states are illusory (Dennett 2018).

Either way, phenomenal consciousness is substituted for the ability to make certain forms of judgement about our representations of ourselves or of the world we perceive. Humphrey described consciousness as an "impossible self," an illusion caused by the way we represent ourselves as existing in the world (Humphrey 2011: 39). Rey argued that it is the functional role that phenomenal concepts play in our overall conceptual framework that characterises consciousness (Rey 1992). Hall supposed that phenomenal concepts are those used by intelligent faculties in the brain as shorthand for computational purposes (Hall 2007). Dennett stated that the apparently qualitative aspects of consciousness are user illusions designed for information processing in the absence of knowledge regarding the underlying processes that make that processing possible (Dennett 1991a: 309–314). Illusionism thus constitutes the belief that consciousness is characterised by our capacity to make a particular form of judgement, although I do not suppose that illusionists necessarily agree on exactly what that form of judgement is. For instance, Frankish has differentiated between illusionism that characterises illusory phenomenal properties as being those belonging to experiences and illusionism that characterises those properties belonging to the objects of our experience (Frankish 2016: 19–20). Either way, I will refer to the forms of judgement that the illusionists posit to account for our belief in phenomenal concepts as "phenomenal judgements".

The question that an illusionist must thus be able to answer is how we can determine what the empirically observable basis for a phenomenal judgement is. This is because illusionism seeks to replace phenomenal consciousness with phenomenal judgements, and thus if the empirically observable correlates of phenomenal judgements are just as difficult to identify through empirical means then illusionism will be victim to an equivalent problem to the

Consciousness Science Paradox. We have some *prima facie* reasons for supposing this should be possible; when I claim that I am seeing two vertical lines as opposed to two horizontal lines, there must presumably be some cognitive capability realised in some physiological mechanism that will correspond to the judgement underpinning this claim. Nonetheless, determining what such capabilities and mechanisms amount to is just the problem that we have already encountered in relation to recent attempts at developing a consciousness science. Specifying that these are not phenomenal properties but are simply those required for phenomenal judgements does nothing to resolve the issue.

To demonstrate, if we wish to determine what the empirically observable correlates of a phenomenal judgement are, we must know what the empirically observable effects of that judgement are. There are cases where we do know at least some of these effects. When a subject truthfully claims to be experiencing some particular quality, this forms a paradigmatic example of the sort of claim being caused at least partially by phenomenal judgements. Such claims are thus examples of recognisable conscious behaviour. This fact alone, though, does not entail that phenomenal judgements are present only where the recognisable conscious behaviour is present. It may be that even where no such behaviour is present there are cognitive or neurophysiological states that correlate with phenomenal judgements.

This was discussed at length already in section [3.2.2] and so I will not repeat the discussion here. Nonetheless, it is worth drawing out the conclusion regarding illusionism now given that the former discussion occurred before the Consciousness Science Paradox had been established.

I will simply observe that it follows that, since such judgements can only be verified provided they are able to produce recognisable conscious behaviour, we would have to assume that they are only present insofar as they can do so. The trouble is that we need some means of distinguishing the presence of true claims about the illusion of phenomenal consciousness from false ones, as we cannot claim to believe that there is an illusion of phenomenal consciousness unless we believe that there are circumstances under which it

would be true to say that the illusion of phenomenal consciousness is present and circumstances under which it would be false to say so. If the properties responsible for my judgement that I am perceiving redness are present, then my claim to be experiencing the illusion of being in the conscious state of perceiving redness is true. Equally, it is possible for me to falsely claim that the properties responsible for that judgement are present, such as if I lie about experiencing the illusion of being in the conscious state of perceiving redness.

It can easily be shown that this ultimately fails to give us the result required by the illusionist. In a situation where one subject is falsely claiming to be experiencing a certain feeling and another is truthfully claiming to be experiencing that feeling, we simply have two sets of conditions, both of which we assume are underpinned by different sets of cognitive or physiological properties. I say "we assume" because it is not the result of empirical observation about the cognitive distinctions between these conditions that has led us to assume that there must be some particular phenomena denoted by the term "the illusion of phenomenal consciousness" that does exist in one set of conditions and not the other, but rather our commitment to this illusion that requires us to accept that one claim is true and the other false. If we were to investigate further, it is possible that empirical investigation will bear out no clear distinction between the two categories. Indeed, even if empirical observation did identify two distinct categories that correspond to our "true" and "false" claims about our experienced feelings, the result of such observations would simply be that we have found two distinct sets of states producing the same set of claims, not that the claims produced by one such set are more imbued with truth or falsehood than those produced by the other. As such, there is no feasible way this would provide us with a means of distinguishing between true and false claims regarding consciousness.

This demonstrates how the Consciousness Science Paradox persists, or that we at least end up with a similar paradoxical result. Whether or not we endorse the claim that consciousness exists, the capacity to judge ourselves to be in a conscious state, which is central to the illusionist's position, should be possible through scientific means and yet

illusionism gives us no provision to perform this study. As such, illusionism is just as problematic as the other aforementioned positions.

[7.5] Summary

In this chapter, we first detailed what the claims specific to neutral monism are by contrasting them with those of panqualityists such as Coleman, before showing how Coleman's view is subject to the Consciousness Science Paradox [7.1]. We then moved on to see how neutral monism, while it could not be said to be subject to the same paradox, has the same paradoxical result with regard to the percepts or experiences it takes to exist in place of consciousness [7.2]. We examined eliminative materialism and showed that it lies between two difficulties; it either fully denies the existence of consciousness, mental content, and even empirical observation, in which case it undermines our knowledge of neurophysiology to the extent that we have no motivation to believe that concepts related to consciousness can be eliminated in favour of neurophysiological concepts, or it has to accept that neurophysiology can underpin empirical observation, in which case it must have a conception of awareness that is just as susceptible to a paradoxical result as physicalism [7.3]. Finally, we explored illusionism where we found that replacing talk of phenomenal states with phenomenal judgements simply means that the paradoxical result occurs in relation to empirically observable correlates of phenomenal judgements rather than phenomenal states [7.4]. In short, we concluded that problems related to the Consciousness Science Paradox, relying on very similar assumptions about possible empirically observable correlations between neurophysiological states and the content of awareness, sensations, percepts, or phenomenal judgements, persist in all of the above cases, and these problems are no easier to solve.

More broadly, over the two previous chapters, we have reached the conclusion that the source of the Consciousness Science Paradox is not a singular metaphysical perspective.

Indeed, the difficulties leading to it are surprisingly prevalent across a varied spectrum of metaphysical positions. Avoiding such difficulties is not something we can do simply by stating that consciousness is separate from the physical, identical to it, or even that consciousness does not exist.

Several of the perspectives outlined posit a relationship between observable states, such as physiological states, and conscious states, whether this relationship is identity, supervenience, or simply co-existence. It does not matter which relationship we endorse; they are all equally capable of falling victim to the paradox. This follows from the fact that the paradox arises because of the joint facts that conscious states must be causally efficacious and that we cannot empirically determine the correlates of conscious states. Even if the relationship is one of identity, the identifications can only be made *a posteriori* and so the paradox persists.

The problem will thus need to be diagnosed. Whatever the problem is, it is somehow common to all of the perspectives mentioned, and yet the difference between the positions described may make such a thing difficult to identify. Nonetheless, I hope to demonstrate that there is an assumption about consciousness that can be attributed to all of the aforementioned positions, even those that deny the existence of consciousness.

8

Diagnosing the Source of the Paradox

[8.1] Observing consciousness

If the Consciousness Science Paradox or some similar difficulty arises in all of the aforementioned positions that defend a certain position on the existence and nature of consciousness, and the relationship between consciousness and the physical, despite their radical disagreement on these matters, there must be something in common between these positions that produces this kind of problem. Even those positions that deny the existence of consciousness seem susceptible to equivalent problems, such that they seem to stem from the same root problem. It should be possible, then, to construct a diagnosis of the problem that can be stated in terms acceptable to both those who assert and those who deny the existence of consciousness. This is something we will address, but it will be simpler to do so after focussing on problems caused for perspectives that maintain the existence of conscious states to discern a common element in these positions that causes the problem.

To ascertain why the paradox should arise, we should attempt to distinguish the common factors that lie across both claims making up the paradox in order to determine where the tension lies. The two most prominent common factors are that both relate to consciousness and to the possibility of identifying conscious states on the basis of empirical observations. On the one hand it is supposed to be possible to correlate conscious states with other events on the basis of empirical observations, but on the other hand this is supposed to be impossible. A reasonable place to look for the source of the paradox, then, lies at how we are to understand the sense in which a conscious state can be empirically identifiable.

We should return here to the question of why consciousness should be empirically *identifiable* rather than empirically *observable*. The justification for this given earlier in the thesis is that even if conscious states were identical to neurophysiological states and thus conscious states could be empirically observed by observing neurophysiological states, we would still have further empirical work to do to determine which conscious states were identical to which neurophysiological states. This issue must now be pursued further. If the work to be done were purely empirical, it would consist of observing conscious states, observing neurophysiological states and then correlating them. Indeed, an *a posteriori* correlative project of some sort has seemed possible in principle even to many who are not reductive physicalists, whether or not they would accept the above statement regarding observations of conscious states. What could constitute an observation of a conscious state in the first place is a difficult question. If we are reductive physicalists, for instance, we might claim that to observe a neurophysiological state *is* to observe a conscious state, yet there must be some other means of observation available otherwise *a posteriori* identifications between neurophysiological states qua conscious states and neurophysiological states qua neurophysiological states would be impossible to make. Identifications of water and H_2O, for example, could only be done following the development of an empirical framework whereby elements such as hydrogen and oxygen were discovered, and our observations of the liquid we call “water” were then to be understood within this overall empirical framework. Without the empirical work required to produce our understanding of molecules, no number of observations of water would have given us any reason to assume that H_2O would be a useful or accurate description of the substance. Similarly, there must be some other empirical means to determine what conscious states are to ascertain which neurophysiological states they are identical to, as no number of observations of neurophysiological states qua neurophysiological states will alone yield this knowledge.

These factors all stem from the fact that empirical correlations can only be made if it is possible to observe a conscious state qua a conscious state in the first place. If I want to make

empirical correlations between the presence of H_2O and water, I must be able to both observe water qua a certain combination of hydrogen and oxygen atoms and observe water qua a certain drinkable, transparent, colourless liquid, and the same applies to conscious states. We need to ascertain, then, how we could observe conscious states qua conscious states. The only two means we could have of observing a conscious state qua a conscious state are either through first-person experience or third-person observations of behavioural indicators. Which of these gives us empirical observations of conscious states qua conscious states?

The multiple problems I have raised with using behavioural indicators to make correlations between conscious states and neurophysiological states have focused largely on the fact that it is impossible to determine that a conscious state is absent in any situation based on observations of behaviour. Nonetheless, I have also focused on PC states, paradigmatic states of consciousness where an individual clearly indicates through their behaviour or reports which conscious state they are in and have assumed that where there is a clear behavioural indicator there is also a conscious state present. Whether this is actually so is a legitimate question (and forms a basis for the Problem of Other Minds), but we need not discuss this here. It will suffice to state that, if conscious states are usually or frequently not present where behavioural indicators are present, then it would be false to state that an observation of a behavioural state constitutes an observation of a conscious state. After all, these behavioural indicators would only sometimes be present where conscious states are, and thus the conscious state would have to cause or otherwise underlie the behavioural state, rather than being identical to it. We would never know whether we were observing a behavioural state caused by a conscious state or not, and thus not only would an observation of a behavioural state fail to suffice as an observation of a conscious state, but we would be unable in such a situation to determine which state we should observe in order to make an observation of a conscious state.

Instead, then, we will assume that behavioural indicators do tell us reliably that PC states are present. The question is whether this suffices as an observation of a PC state.

What we need to know here is whether an observation of a behavioural indicator is an observation of an effect of a PC state or a direct observation of a PC state itself, as these are very different claims. First, I will address the possibility that an observation of a behavioural indicator is an observation of a PC state.

This possibility requires us to assume not only that behaviourism is true but that the dispositional states required in our formulation of behaviourism can be observed directly via observation of instances of behaviour. It is simple to demonstrate that this cannot be so. If pain was a tendency to behave in a certain way, such as to scream whenever a hammer is dropped on your foot, and a tendency to behave a certain way was nothing over and above certain behaviour (i.e. screaming whenever a hammer is dropped on your foot) then you would not be in pain if you failed to scream, even if you failed to do so because you were suppressing the urge to scream in order to avoid waking somebody or even because you were injected with muscle relaxants and could not move or make any noise. I will assume that any analysis of pain that assumes that it is not present where there is not a single, specifiable behaviour is inadequate. In any case the behaviourist's claim would usually be that you would be in pain in such a situation but due to either physical circumstances (i.e. being injected with muscle relaxants) or the presence of another behavioural disposition (i.e. the disposition to suppress your screams) the behaviour was not carried out. If this is the case then the behavioural disposition is not observable; you may see a certain behaviour but you will have to derive which dispositions are being acted out from further observations of that person's behaviour, assuming that this could ever give you the final story about which dispositions are present. In this case, the situation is indistinguishable from one where the presence of a conscious state is derived from behaviour rather than being directly observed.

The more plausible situation is thus the one in which observations of behavioural indicators allow us to derive the presence of PC states. That is, we assume that the behaviour is present because the conscious state is present rather than that the behaviour itself is identical to the conscious state. There are plenty of examples in science of entities or

processes that are inferred from others that are observed. However, such theoretical entities are posited because of their role in accounting for empirically observed facts. For instance, dark matter may not be directly observable, but one reason that it has been posited is to account for our models of galaxies containing too little visible matter for them having the structure that they do (Trimble 1987), and the effects of the presence of dark matter are assumed to be observable through gravitational lensing, where an image of distant galaxies is distorted by intervening mass bending the path of light transmitted from those galaxies (Natarajan et al. 2017). Dark matter is thus posited to fill a gap between our estimates of visible matter in galaxies and our estimates of the overall quantity of matter in galaxies, with consequences regarding what we predict to observe in certain situations.

Consciousness, though, plays no such role. I could in principle explain a person's behaviour by only positing cognitive processes, certain states of information, physiological states, and historical causal processes. There is no element of human behaviour, nor information-processing capabilities, that require us to posit consciousness in order to explain it. In fact, the only explanatory framework regarding the nature of human beings or the world we inhabit that requires us to posit consciousness is that of our understanding of consciousness itself. In explaining all other characteristics of humans in purely cognitive, information-processing and physiological terms, I would be developing explanations without anywhere near the level of deep ambiguity in supposedly empirically confirmed claims about consciousness. This is in itself an argument against those eliminativist projects that claim that consciousness exists but is nothing over and above certain cognitive or physiological states, and that phenomenal qualities do not exist. If we wish to describe human behaviour in such terms, then what is the purpose of calling any of it "consciousness?" It may indeed be possible to revise the concept of consciousness such that it does not include any problematic phenomena, but if the project is simply to describe human behaviour in such terms, this would be equally possible by eliminating consciousness entirely and talking only of cognitive and physiological states. Consciousness is thus not a theoretical necessity in such an explanation.

Indeed, the project of rehabilitating consciousness within this role distracts from the purpose of explaining human behaviour in such terms; it might have been possible to describe the soul in purely physical terms with enough conceptual development and further philosophical work but doing so would surely be nothing more than a distraction for somebody who wished only to understand the workings of the human body.

A more likely reason we derive the existence of PC states in others is not that there is something about human behaviour that seems *in principle* impossible to explain unless we posit them, but rather that we deem ourselves to behave in the ways that we do because of our conscious states. Just as I smile and laugh when I am feeling happy, I judge others to feel the same when they behave in the same way. If this is so, then it is only in our first-person perspective that we empirically identify the presence of conscious states, and as a result of this we attribute conscious states to others who have similar behaviours. This is the possibility we will now turn to.

We may think that the clearest evidence we have for consciousness is our own first-person experience. The states apparent to me in my subjective perspective on the world are, by our previous definition, conscious states:

Conscious state: An instance of subjective perspective.

The question we are interested in addressing here is whether we *observe* a conscious state in having a subjective perspective.

It seems that whichever way we address this question, we cannot arrive at a description whereby it makes sense to say that we observe our own subjective perspective in the same way that we observe states of the world lying outside of our consciousness. Observing a tree consists of having a certain subjective perspective, so it cannot be true that I observe this perspective in the same way that I observe the tree.

This is an issue that has been brought up many times in discussion over the "openness" of experience. It is not as though when we have a perception we encounter two sets of states, one set belonging to the world we are perceiving and one set belonging to our perspective. Valberg wrote about a "horizonal" conception of experience, whereby we take the totality of our experience and, within this totality, we are unable to discern any particular character of experience (Valberg 1992: 124–125). This is because experience is not a part of the world but rather "it is a 'limit' of the world" (ibid.: 125). This conception Valberg drew from Wittgenstein, who stated, "The subject does not belong to the world: rather, it is a limit of the world" (Wittgenstein 1921: 5.632). Wittgenstein argued this on the basis that just as the visual field does not have the eye that sees within it, our experience does not have the experiencing subject within it. Valberg's argument is that if somebody describes what an object does, such as keeping tools in it, we can ask the further question, "Yes, but what *is* it?" In such a situation, we want to know what it is in itself even if was not being used to keep tools in it. However, with experience, if we state that it is that within which certain things are present to us, the further question, "Yes, but what *is* it?" "has no answer" (Valberg 1992: 123). There is nothing about the character of the experience that can be distinguished from the presence of an experienced thing such that we can state that *this* is the character of our experience.

The indistinguishability of the character of experience and that of the objects of experience forms the ground for discussion of the transparency of experience. Tye made the point that experience is transparent:

> If you are attending to how things *look* to you, as opposed to how they are independent of how they look, you are bringing to bear your faculty of introspection. But in so doing, you are not aware of any inner object or thing. The only objects of which you are aware are the external ones making up the scene before your eyes. (Tye 2000: 46–47)

We could mistake the above passage to mean that in veridical perception the qualitative properties we attribute to consciousness actually belong to entities lying outside of our awareness but Tye emphasised that something only gains phenomenal character when it enters into representational content (ibid.: 48–49). In the above passage, his point was that we do not bring to mind a distinct set of phenomenal qualities when we reflect on the character of our perceptual awareness; the properties we are aware of simply have that phenomenal character. When we attempt to introspect on the character of our perceptual states, we simply bring to mind that which we perceive.

If we accept transparency regarding experience, then we do not observe our subjective perspective in the same way that we observe trees, buildings or stars. We may think, then, that we can avoid the issue if we reject transparency. Yet, the issue is the same even if we totally disagree with the above analysis and hold instead an opacity view of experience whereby we are only indirectly aware of states of the world outside of our consciousness via those that appear directly to us within our consciousness. In such a view, we are only ever aware of conscious states even in veridical perception, and if this is so then conscious states are observable in a way that other objectively existing states of the world are not. Either way, our subjective perspective cannot be described as observable in the same way that the location of an object that can be independently confirmed by multiple subjects is observable.

The same discrepancy between empirical knowledge and knowledge of conscious states is apparent in epistemological perspectives that argue that we are entitled to greater certainty regarding conscious states than we are to empirically observable states. Goff described our knowledge of our conscious states as being “super-justified” (Goff 2017: 112). He claimed that, while we can be wrong about our conscious states, claims about our conscious states are less susceptible to scepticism than empirical beliefs. Yet, if this is so, this implies that the process by which we know about our conscious states is distinct to that which we know about other empirically observable states, and thus the two are not observable in the same sense at all.

The reason that this discrepancy between our knowledge of consciousness and empirically observable states of affairs should cause a paradox can now be drawn out. The conditions required for us to gain awareness of the presence of conscious states are different from those required for us to gain awareness of the presence of other states. It may appear in a sense that this is not so; to take, for instance, my awareness of a table, I can be aware that the table is there and is in a certain state by being in a certain perceptual state, and to be aware of my perceptual state I simply need to be in that very state. Yet, for another person to become aware of the numerically same state of the table, they need to be in their own perceptual state. There is no corresponding way for them to become aware of my perceptual state, unless I reveal the presence of that state through my behaviour or some form of communication. The difficulties in this manner of ascertaining the presence of conscious states have been explained in detail already; we have to make numerous assumptions regarding the relationship between the presence of conscious states and behaviour if we are to discern whether conscious states are present even where such behaviours are absent. This makes it impossible to make clams about which conditions are required for conscious states to be present.

McDowell might have retorted that such cases do not require us to infer a conscious state at all. For instance, he stated that we can construe a theory of realism whereby, by learning the truth conditions for the presence of pain, "one can literally perceive, in another person's facial expression or his behaviour, that he is in pain, and not just infer that he is in pain from what one perceives" (McDowell 1978: 136). This sort of statement has implications regarding cases whereby we might claim that somebody behaving as though they are in pain may not be, but it has no implications regarding cases whereby we might claim that somebody is in pain despite producing no behavioural indication that they are. However, what is meant here is not that pain is identical to pulling a certain kind of facial expression, but that the presence of a certain kind of facial expression is a truth condition of the presence of pain. As

such, pain can be present without such a facial expression, and that would not contradict McDowell's point. As McDowell put it:

> What we have to deal with, primarily, is *general* competence with other-ascriptions of pain… the realist cannot claim that each such sub-competence, if described as involving a conception of a truth condition, is directly manifestable in behaviour… Ascription of the general competence is justified if events construable as manifestations of the implied sub-competences actually present themselves to observation in the case of *some* utterances of the relevant sort... (ibid.: 139; emphasis in the original text)

The problem then is essentially the same. Whether or not we can, in certain cases, confirm the presence of a conscious state because of the presence of certain behaviour, we are still in no position to contrast the presence and absence of conscious states, which was what produced the paradox in the first place.

What we have ascertained so far, then, is an exceptional feature of conscious states. They cannot be identified on the basis of empirical observations. An individual does not observe a conscious state, they are aware of it simply by being in it. It is not surprising, then, that empirical correlations cannot be made because our knowledge of conscious states does not come from empirical observation; we cannot correlate one thing that we know about through empirical observation with another thing that we know about but cannot empirically observe. An individual identifies the presence of a conscious state in any instance by being consciously aware of something.

All true statements about conscious states can take the form, "X is/was consciously aware of Y." The truth of such a statement cannot be ascertained by anybody other than the individual in the conscious state, or else it must be derived from that individual's behaviour. There are two ways we can understand this, depending on whether we accept that X can be

consciously aware of Y even where there is no intersubjectively identifiable instance of a Y present. An example of this would be if I could be consciously aware of the presence of a computer even while there is no computer actually in front of me.

If we accept that this is so, this entails that the "Y" in the above form of statement cannot be empirically identified by anybody other than the individual in the conscious state. At this point we have already diverged from the sort of description that could be found in relational statements the truth of which can be evaluated as a result of empirical observation.

For instance, I can state that, "X has written Y," or that, "X is standing in front of Y," in which case whatever the X in question is, we can empirically determine that it is indeed the case that it has written whatever the Y in question is or that it is standing in front of whatever the Y in question is. Moreover, where X stands for an empirically identifiable entity it cannot write something that is *in principle* unobservable, nor can it stand in front of something that is *in principle* unobservable. However, the Y in question as an object of conscious awareness cannot be similarly empirically verified or refuted.

Since, then, observations of the presence or absence of Y do not tell us whether this conscious state is present, the only other aspect that can be empirically identified is the presence or absence of X, which in this case would be the experiencing subject. Yet, there is no way to identify the presence of an experiencing subject without identifying them being in a particular experiencing state. As such, since the Y being experienced cannot be empirically verified through identifying the presence of whatever Y is an experience *of*, we equally cannot empirically identify the presence of an X that is experiencing Y. The only way that we can empirically identify the presence of X or Y is by *being* X and being consciously aware of Y ourselves.

The diagnosis, then, is this:

> **Paradox Diagnosis:** PC states must be empirically observable because they can produce reliable, true claims about themselves, but statements of the kind, "X is aware of Y," cannot be confirmed or refuted via empirical observation.

If we instead reject that X being consciously aware of Y entails that there is no intersubjectively observable Y present, we may appear to have a position that avoids this diagnosis. An example of a perspective that would reject such a thing is naive realism.

[8.2] Naive realism and disjunctivism

Naive realism is, broadly, the thesis that in veridical perception we are directly acquainted with the actual properties of the objects in the physical world that we perceive. It has been proposed as an alternative to other prominent views of perception by several authors, such as McDowell, Snowdon, Martin and Brewer.

McDowell argued that, while we may seem to perceive things even while those things are not there, our perceptual experiences only have the character they do as a result of them being suitably caused by our environment. He described the relation between perceptual knowledge and the objects of this knowledge in this way:

> Perception makes knowledge about things available by placing them in view for us. But it is precisely by virtue of having content as they do that perceptual experiences put us in such relations to things. (McDowell 2013: 144)

This may not appear to directly imply naive realism, but McDowell stated further that the fact that experiences have content does not imply "denying that if an experience is one of seeing, what its subject encounters in her experience is an environmental reality" (ibid.: 146).

If we do not deny this, what is encountered in perception is not something that belongs to conscious states from which external realities are derived but are rather themselves aspects of the environment that the subject subsequently has knowledge about.

For McDowell, the difference between perceptions and cases where we only *seem* to have a perception, such as when we seem to see an object but there was no such object appropriately placed to have caused that perception, can be described by virtue of how our experiences have the content that they do (ibid.: 147). He added further that "the experience's epistemic significance must be part of its subjective character" (ibid.: 149). So, if we perceive something, it is because we have a certain representational content, which we have by virtue of our environment producing in us the perception that that self-same environment is present, and this situation is qualitatively distinct from one in which we merely seem to perceive that environment. To put the point somewhat concisely:

> One's knowledge that there is something red and rectangular in front of one includes knowledge of its own credentials as knowledge. And it is the knowledge it is because it is a non-defective act of a capacity to know such things through perception. (ibid.: 151)

When we know something through perception, according to McDowell, an environmental reality is present to us (ibid.:151).

Snowdon similarly argued that naive realism is a radical alternative to the theory of perception captured in what he called "the causalist viewpoint" (Snowdon 1981: 175). The causalist viewpoint includes the theses that an object must have a causal impact on a subject for that subject to perceive it and that objects produce states in subjects that can be described in statements such as, "It looks to *S* as if…" and that these words can be interpreted "phenomenologically (rather than as ascribing, say, a tentative judgement by *S*)" (ibid.: 176).

To outline his alternative, Snowdon referred to work by Hinton wherein he argued for such words to be interpreted not phenomenologically, but disjunctively (ibid.: 184).

Hinton argued that statements such as, “It looks to *S* as if there is an O” should be interpreted as meaning that either S perceives an O or else S only seems to perceive an O. That is, there is no reason to suppose that the subjective indistinguishability between veridical perception and illusion can only be explained by reference to there being some “specific and well-defined type of event” in common between two such cases (Hinton 1973: 77). He used the example of somebody having a certain visual experience that they could describe as being either that of seeing a flash of light or having the perfect illusion of doing so (ibid.: 39). Disjunctivism constitutes a rejection of the claim that there is a particular visual experience that fits both of these circumstances, supposing instead that an illusion of seeming to see a flash of light and a perception of seeing a flash of light fit into a generic category of which both states could fit, and are not a distinctive kind of event of their own (ibid.: 80).

Snowdon argued that demonstrative judgements, such as that “that is a light bulb,” can be made sense of either via a causalist viewpoint, whereby we encounter a phenomena that we could be aware of both in cases where our judgement is true and in cases where it is false, or by supposing that the phenomenon we refer to in accurate judgements of the presence of a light bulb is only present in such correct judgements, and as such is only present in cases where there is a light bulb situated so as to allow an accurate demonstrative judgement to be made about it (Snowdon 1981: 187). He later went on to argue that “the fact that two cases are indistinguishable to the subject does not mean that they are not, in themselves, quite different” (Snowdon 2005: 287). He opted instead for the position that when a state that seems to be a perceptual state is produced, there may instead be an “inner experience,” such that its “intrinsic character and nature is independent of any item in the world spatially external to the subject’s body” (ibid.: 288), and that in cases of hallucination, for instance, such experiences can be distinguished from the experience of having a veridical perception (ibid.: 303).

Brewer, in arguing for his version of naive realism, which he called “the object view” (Brewer 2008: 171), said rather that things that one is experiencing in hallucinations are visually indistinguishable from things that one is experiencing in veridical perception and that “no more positive characterization of the experience may be given” (ibid.: 173). This is a view he attributed to Martin, who claimed that we should not be able to distinguish between a hallucination and perception even if both situations are distinct in kind, since this supposes that “a responsible subject who wishes to determine how things are with him or herself through reflection must not only correctly identify phenomenal properties of a specific sort when they are present, but also they cannot be misled into judging them present when they are not” (Martin 2004: 50–51). This, Martin argued, is problematic because it is at least questionable that we should be infallible in our judgements regarding our sensory experiences given that we can be misled about what type of sensory experience we are having (i.e. hallucination or perception) (ibid.: 51). It could perhaps be stated that to argue for this infallibility in our sensory judgements is to beg the question against the naive realist.

It should be noted that elsewhere Martin suggested the alternative that in non-perceptual experiences we may “deny that there is any object of experience at all” in cases such as hallucinations (Martin 2002: 395). While Martin similarly argued for naive realism, he did so as an alternative to intentional theories of perception, such as that held by Tye, which posit that the qualities we perceive belong to our representational content and this content can be present whether it is caused by the presence of the phenomena we believe ourselves to be perceiving or not. Under intentional theories, we might hallucinate that we are perceiving the Pacific Ocean, for instance, and thus represent blueness (ibid.: 384). Instead, we might follow in Martin suggesting that we perceive blueness in cases of veridical perception and in hallucination it might be better to deny that there is any blueness present at all.

This gives us three different ways of accounting for illusions and hallucinations. Either:

(A) Hallucinations and illusions involve the presence of mind-dependent qualities of experience.

(B) Hallucinations and illusions involve the absence of all qualities of experience and instead involve only judgements regarding qualities of experience.

(C) We can say regarding hallucinations and illusions only that we are unable to subjectively distinguish them from veridical perceptions and nothing more about their character.

All three of these possibilities make cases of hallucination and illusion as paradoxical as they were under those views that do not accept naive realism.

If (A) is true, then we have to account for the presence of hallucinations and illusions by accounting for the presence of qualities of experience present only to the subject that has them. This means that we cannot determine whether or not a quality of such experience is absent unless a subject claims that it is absent, and even so we would then have to assume that the presence of such qualities entails the presence of recognisable conscious behaviour.

If (B) is true, then we have to account for the presence of hallucinations and illusions by accounting for the presence of judgements of qualities of experience present only to the subject that has them. This is the same problem suffered by illusionists; individuals only judge themselves to be in certain conscious states given the presence of certain states, but these states can only be specified if we can contrast the presence of such judgements with their absence. To do this, we must assume that judgements about consciousness are present only where recognisable conscious behaviours are present.

If (C) is true, then there are circumstances under which it is indistinguishable to one whether or not they are having a perception. These circumstances can only be known by the subject that has them. This means that we cannot determine whether or not these circumstances are present unless a subject claims that it is absent, and even so we would

have to assume that the presence of such circumstances entails the presence of recognisable conscious behaviour.

Once more, we cannot justify our assumption that a conscious state is only present where there is recognisable conscious behaviour on the basis of empirical observations. The naive realist could retort to these criticisms that they fail to take into account that there are certain things we can know with regard to veridical perceptions. For instance, we could know that in cases of a veridical perception of, for instance, a red and rectangular shape, such-and-such perceptual apparatus is functioning in such-and-such way and, as such, we can know that if the same perceptual apparatus is functioning in a similar way without the actual presence of a rectangular shape, we would be justified in assuming that what is present is a mere seeming-to-be-seeing a red and rectangular shape rather than an actual seeing of a red and rectangular shape. There is, though, a bigger problem with such an analysis that implicates even the naive realist's claims about ordinary cases of perception in paradox.

This problem stems from the fact that it must be by virtue of the presence of certain states that it can seem to a subject that they are perceiving something. While the disjunctive thesis implies that we experience something different in perception to what we experience in cases of hallucination and illusion, what lies in common between the two and grounds both cases in the disjunction of perceiving and merely seeming to perceive is some state that allows us to form the judgement that we are perceiving something. While there might be some aspect of your overall neurophysiological or cognitive state in having a veridical perception that can account for our capacity to form such judgements, the only way we can determine what it is is by contrasting the presence of them with their absence. Since this cannot be done without assuming that their presence must produce recognisable conscious behaviour, we have no way of distinguishing the presence of such judgements from their absence. This means that we cannot know which parts of our neurophysiological or cognitive apparatus are essential for our judgements as opposed to simply being part of an information-processing capacity that would, in isolation, produce no judgement about the presence of a perception at all.

While we can, then, determine that a veridical perception is not present through independent means, such as by observing that there is in fact no object suitably placed to you to warrant your claim to be perceiving it, we equally cannot determine that one is present other than via a subject's first-person perspective and subsequent behaviour.

One account of naive realism that does not rely on the disjunctive thesis was presented by Conduct, who stated that naive realists could instead assume that in veridical perception we are presented with universals instantiated by particulars, whereas in hallucination we are presented only with universals (Conduct 2012: 732). Nonetheless, this position is just as susceptible to the paradox: making awareness of a universal the common element between perceptions and hallucinations simply means that awareness of universals is just as susceptible to the paradox, since we can only know that such awareness is present from the first person or from inferring such where certain behaviours are present. I will thus suppose that this perspective fails to resolve the issue.

For the naive realist, consciousness plays the same role as it does for other theorists; conscious states can only be identified indirectly through empirical observations of somebody's behaviour, which they perform as a result of their first-person experience. This is because both disjunctivism and awareness of universals, either of which is required for naive realism to make sense of illusions and hallucinations, requires that it can seem to one that they are having a perception whether they are or not, and this "seeming-to" is a common element between both perceptions and non-perceptual experiences. It may be so that the two kinds of state are fundamentally different in terms of what is presented, but for it to be a state of either kind it still must seem to the individual that they are having a perception.

It may well be that in cases where an individual behaves as though they are in a conscious state we empirically confirm that they are by observing that behaviour, but since this does not tell us whether the behaviour is essential for the presence of the conscious state, the only means we have of ruling out the presence of a conscious state is through knowing what is in the first-person perspective of all relevant subjects. Naive realism thus still gives

consciousness a role whereby the presence of its states is only knowable indirectly via the first-person perspective of subjects.

It does, however, admit a slightly different analysis to other positions, but only by pushing the analysis back one step. "X is aware of Y" is possible to refute by observing that no Y is present, if Y stands for a mind-independent object that the statement claims X to be perceiving, and so statements of this kind are empirically refutable in a way that they are not under other positions. Nonetheless, given disjunctivism, what is required for the subject to be in either a perceptual or non-perceptual conscious state is for it to seem to that subject that they are perceiving Y, although the conscious state in question differs whether they actually are perceiving Y.

As such, the common element in the disjunction is X seeming to be aware of Y, and this element, critical to both perceptions and non-perceptual experiences, admits the same analysis as "X is aware of Y" does in the other positions mentioned. "X seems to be aware of Y" is not possible to refute by observing that no Y is present; I may seem to perceive a tree even if no tree is present. The same role is being performed by a subject "seeming-to" perceive something as the content of conscious awareness plays in the views that naive realism opposes. There may be multiple kinds of conscious state involved between perceptual states and states where an individual is mistaken about being in a perceptual state, but all of them require it to be the case that a subject seems to perceive something, and since it is in principle impossible to refute empirically that this is taking place, the empirically observable correlates of conscious states cannot be identified.

The paradox diagnosis for naive realism is as follows:

> **Paradox Diagnosis 2:** PC states must be empirically observable because they can produce reliable, true claims about themselves, but statements of the kind, "X seems to be aware of Y," cannot be confirmed or refuted via empirical observation.

This diagnosis is superior to the previous one because it works both for naive realism and for the positions captured by the previous diagnosis. Even if we reject naive realism and claim that we are in conscious states because we are aware of something, it still follows that we are in that conscious state because we seem to be aware of something; it just happens that seeming to be aware of a thing is an identical psychological state to being aware of that thing.

Naive realism fails to avoid the paradox and falls victim to it because it admits a critical element in our judgements about our conscious states that is analysable in the same terms as those views it stands in opposition to.

[8.3] Diagnosing positions that deny the existence of consciousness

The problem that has been diagnosed so far in the analysis is that statements of the kind, "X seems to be aware of Y," are doomed to admit no possibility of empirical confirmation that is not indirectly produced by the first-person awareness of a subject. We might wonder once again why we cannot simply avoid this problem by denying the existence of consciousness altogether and whether this would avoid the above sort of diagnosis.

It is worth recapping what the problem with such a perspective was in the first place. The problem was that eliminativist perspectives cannot deny the existence of empirical knowledge without this undermining any basis for supposing that we can describe anything in terms of neurophysiological or cognitive states at all. Unfortunately, if we allow for the existence of empirical knowledge, we are endowed with instances of awareness that do not avoid the same problems as those admitted by the individual who believes in the existence of consciousness.

If we take illusionism as such an example, all statements of the kind, “X is aware of Y,” should be replaced by statements such as, “X judges her/himself to be aware of Y.” This is because the apparent existence of phenomenal qualities is, according to the illusionist, better explained as simply being the existence of a state whereby an individual judges her/himself to be experiencing a phenomenal quality, and so it is only the judgement that requires explanation. Nonetheless, this simply pushes our analysis back one step again; “X judges her/himself to be aware of Y” is something that can only be established by the individual making the judgement or indirectly by others who observe the behaviour of that individual. The first-person perspective is still given unique epistemic authority in such cases such that a third-person observation can at best only let us infer that such a state is present.

If we instead attempt to describe “X judges” in terms of specific observable states, such as that somebody only ever perceives Y if they are in a specific neurophysiological state, the work would still need to be done to identify these neurophysiological states. These identifications could not be performed through correlating judgements as identified from the first person with neurophysiological states as identified from the third person as this would produce the same paradoxical result; we would only be able to identify those judgements from the first person that produce recognisable conscious behaviours. The trouble is that there is no way of simply reflecting on our own neurophysiological states “from the inside” to ascertain which are essential for judgements; we would have to be able to make judgements about states that we could empirically observe by virtue of being in neurophysiological states. In other words, I could only determine that a judgement requires a certain neurophysiological state if by being in a certain other neurophysiological state I could make judgements of my own.

I cannot, then, avoid statements of the kind, “X judges her/himself to be aware of Y,” without undermining the entire basis for empirical knowledge in the first place. Of course, as discussed before, the option is open to the eliminativist to bite the bullet and deny the possibility of empirical knowledge, but this risks collapsing eliminativism into a sophisticated version of my parody bean theory, arguing as it does that all states can be described in terms of

neurophysiological states while rendering itself incapable of describing what the benefits of this would be or how such a thing would even be possible.

The paradox diagnosis for eliminativism is as follows:

> **Paradox Diagnosis 3:** PC states must be empirically observable because they can produce reliable, true claims about themselves, but statements of the kind, "X judges her/himself to be aware of Y," cannot be confirmed or refuted via empirical observation.

Denial of the existence of consciousness is thus, on its own, inadequate. A perspective that requires statements of the kind, "X is aware of Y" or "X judges her/himself to be aware of Y" in order to have any explanatory use has the same paradoxical grounding as any rival perspective. They have denied the existence of consciousness, but they have not demonstrated any ability to deny the requirement for statements that perform the same role as statements about conscious awareness, and thus they simply end up transcribing the same paradox into the eliminativist's language. The problem cannot be avoided by simply denying the existence of consciousness; an alternative framework would need to be built that did not require the use of such statements.

Paradox Diagnosis 3 is superior to Paradox Diagnosis 2 because all cases where an individual seems to be aware of something are also cases where an individual judges her/himself to be aware of something. Thus, this diagnosis encompases all cases of eliminativism as well as naive realism and other views that assert the existence of consciousness.

This diagnosis gives us a final conclusion. The paradox is inherent in the very idea of consciousness. All conceptions of consciousness, including those that deny its existence, assume that judgements regarding a subject being in a particular conscious state are possible. This entails that the paradox is a difficulty for all philosophical positions that have something

to say about consciousness, and it is thus one that any philosopher adhering to such a position and wishing to be able to present and defend theories without contradiction should have the desire to resolve.

[8.4] Summary

In this chapter, we have attempted to find what it is about our understanding of consciousness that has produced the paradoxical result that is so prevalent in our understanding of consciousness. Our first avenue into this investigation was to explore the idea of observing conscious states, where we realised that the very idea of observing a conscious state seems flawed [8.1]. This gave us a diagnosis of the paradox whereby it is because statements regarding somebody being consciously aware of something cannot even be confirmed by observation even in PC states. We attempted to avoid this diagnosis by looking at naive realist perspectives on perception to find cases whereby the presence of elements of consciousness can be ascertained through empirical observation because they lie in a publicly observable place, but ultimately the paradox persisted for disjunctive claims regarding a subject seeming-to be in a conscious state [8.2]. We then extended our diagnosis to cases where consciousness is denied to establish that a similar diagnosis can be applied to statements regarding a subject simply judging themselves to be in a certain conscious state [8.3]. This left us with our final diagnosis being that statements regarding a subject judging themselves to be in a certain state produce a paradoxical result.

We will now move on to some concluding comments to reassess the distance we have covered, make some suggestions about further work to be done, and draw out the implications of this paradox on wider philosophical work.

9

Concluding Comments

[9.1] The journey so far

I have attempted to judge how the concept of consciousness fares in the context of a scientific study. To make this judgement as far-reaching as possible, I gave a very inclusive definition of consciousness, as being the capacity to have subjective perspective, without defining what this should imply with regard to whether consciousness is physical or nonphysical, or what this should imply with regard to whether or not consciousness is inherently qualitative or. In order to make this judgement as unambiguous as possible, I have attempted only to assess the kinds of conscious state that produce clear evidence of their existence, which I termed, “paradigmatic states of consciousness,” or “PC states.”

These states were found to meet two contradicting criteria. They were at once found to be causally efficacious, able to produce empirically observable effects, such as behaviours and speech patterns. On the other hand, they were found to be impossible to empirically identify. The efforts of scientists to determine what the empirically observable correlates of such conscious states were in vain; such scientists were forced to make implicit metaphysical assumptions that are not required in other areas of scientific research, and any philosophical grounding that such assumptions might seek, such as through the verificationist tradition, were found not to support them at all.

This paradoxical result could not be avoided through switching from one philosophical view on the nature of science to another, nor could it be avoided through switching from one metaphysical view on the nature of consciousness to another. While an exhaustive analysis

of all philosophical positions on these subjects was not possible, it was at least made clear that the paradox is a difficulty that many popular positions must contend with.

From here, we began to diagnose the cause of the problem and found that all statements regarding a subject's judgement to be in a particular conscious state inevitably produce the same paradox. The final destination of this thesis was to realise that the concept of consciousness is paradoxical in a way that should affect all positions that have something to say about consciousness.

[9.2] What's next?

It is reasonable to wonder what the purpose of pointing out the paradox is, and what we could hope to achieve by finding a resolution. This supposes that a resolution is even possible and, while I am optimistic about this, I have offered no argument to the effect that it should be.

It may even be the case that certain perspectives that already exist offer us a means of avoiding it. I have yet to find such a perspective, but there are entire traditions of thought that I have not had the space to offer an analysis of here. Idealism offers a perspective on the nature of the mind and the nature of the world that radically differs from any of the positions examined in this thesis, and so it would be presumptuous of me to boldly state without argument that no means of avoiding or resolving the paradox lies in this area of thought. There are also differing conceptions of mind and world to come from distinct cultures of philosophy, such as those emerging from Buddhist and other Indian thought, and both a lack of space and a profound ignorance have prevented any analysis of such thought and how the paradox outlined here would fare in their light.

As for the areas discussed, it is difficult to say how consciousness researchers should conduct their studies differently given the existence of the paradox. We may take an entirely

destructive approach and assume that the paradox demonstrates their positions to be hopelessly flawed. We may take a more constructive approach and assume that there are specific ways such positions could be amended to avoid the paradoxical result. We may even assume that the paradox is so prevalent that avoiding it is not a priority.

Of course, any of these conclusions would require it to be widely accepted that such a paradox exists and much work is to be done if the existence of the paradox is to receive such acceptance.

It is very difficult to speculate on what a resolution to the paradox, if one is possible, would look like. Perhaps mostly readily imaginable is that one of the claims will either turn out to be based on flawed reasoning or that our conception of consciousness or science will be modifiable in some way that I have not explored here such that we will be able to conceive of consciousness as either being possible for science to study or as having a nature such that empirical investigation should not be expected to be able to identify it. Less readily imaginable, although the option that I prefer, is that the cause of the paradox is that the concept of consciousness is a composite concept, representing multiple different things, and that once these are considered separately no single concept will both be found to refer to something both causally efficacious and out of the bounds of empirical investigation. This is an avenue I wish to explore in greater detail but have not had the space to do here and so no more will presently be said about it.

In short, there are several different outcomes possible, but none of them are likely until the paradox as it has been presented here is acknowledged as an issue by other philosophers.

REFERENCES

Aicardi, C. (2016) Francis Crick, Cross-Worlds Influencer: A Narrative Model to Historicize Big Bioscience. *Studies in History and Philosophy of Biological and Biomedical Sciences*, 55: 83–95.

Allahverdyan, A.E., Balian, R. & Nieuwenhuizen, T.M. (2012) Understanding Quantum Measurement from the Solution of Dynamical Models. *Physics Reports*, 525(1): 1–166.

Alter, T. & Coleman, S. (Forthcoming) Panpsychism and Russellian Monism. In Seager, W. *The Routledge Handbook of Panpsychism*. London: Routledge.

Ayer, A.J. (1936) *Language, Truth and Logic.* London: Penguin Books.

Ayer, A.J. (1977) The Causal Theory of Perception. *Proceedings of the Aristotelian Society, Supplementary Volumes,* 51: 105–125.

Baars, B. (1988) *A Cognitive Theory of Consciousness*. Cambridge: Cambridge University Press.

Baars, B. (1994) Contrastive Phenomenology: A Thoroughly Empirical Approach to Consciousness. *Psyche*, 1(2): 32–55.

Baars, B. & Franklin, S. (2003) How Conscious Experience and Working Memory Interact. *Trends in Cognitive Sciences*, 7(4): 166–172.

Bachmann, T. & Francis, G. (2014) *Visual Masking: Studying Perception, Attention, and Consciousness*. Oxford: Academic Press.

Batthyány, A. & Elitzur, A. (2009) *Irreducibly Conscious: Selected Papers on Consciousness*. Heidelberg: Universitätsverlag Winter.

Bauer, G., Gerstenbrand, F. & Rumpl, E. (1979) Varieties of the Locked-in Syndrome. *Journal of Neurology*, 221: 77–91.

Bayes, T (1763) An Essay Towards Solving a Problem in the Doctrine of Chances. Communicated by Price, R. in a letter to Canton, J. *Philosophical Transactions of the Royal Society*, 53: 370–418.

Bayne, T. (2010) *The Unity of Consciousness*. Oxford: Oxford University Press.

Becker, E. (1973) *The Denial of Death*. New York: Free Press.

Bennett, J. (1971) *Locke, Berkeley, Hume: Central Themes*. Oxford: Oxford University Press.

Berkeley, G. (1710) *A Treatise Concerning The Principles of Human Knowledge*. Reprinted in Winkler, K. (ed.) (1982). Indianapolis: Hackett Publishing Company.

Bhaskar, R. (1975) Feyerabend and Bachelard: Two Philosophies of Science. *New Left Review*, 94: 31–55.

Bigelow, J.C. & Pargetter, R. (1990) Acquaintance with Qualia. *Theoria*, 61(3): 129–147.

Blakeslee, S. & Ramachandran, V.S. (2005) *Phantoms in the Brain.* New York: Harper Perennial.

Block, N. (1978) Troubles With Functionalism. *Minnesota Studies in the Philosophy of Science*, 9: 261–325.

Block, N. (1995) On a Confusion About a Function of Consciousness. *Behavioural and Brain Sciences*, 18: 227–287.

Bolender, J. (2001) An Argument for Idealism. *Journal of Consciousness Studies*, 8(4): 37–61.

Bor, D. (2012) *The Ravenous Brain*. New York: Basic Books.

Brewer, B. (2008) How to Account for Illusion. In Haddock, A. and Macpherson, F. (2008) *Disjunctivism: Perception, Action, Knowledge.* Oxford: Oxford University Press.

Campion, J., Latto, R. & Smith, Y.M. (1983) Is Blindsight An Effect of Scattered Light, Spared Cortex, and Near-Threshold Vision? *Behavioural and Brain Science*, 6: 423–486.

Carnap, R. (1928) *Der Logische Aufbau der Welt*. Reprinted in George, R.A. (translator) (2003) *The Logical Structure of the World and Psuedoproblems in Philosophy*. Illinois: Open Court.

Carnap, R. (1936) Testability and Meaning. *Philosophy of Science*, 3(4): 419–471.

Chalmers, D. (1995) Facing up to the Problem of Consciousness. Reprinted in Chalmers, D. (ed.) (2010) *The Character of Consciousness*. 3–28. Oxford: Oxford University Press.

Chalmers, D. (1996) *The Conscious Mind*. Oxford: Oxford University Press.

Chalmers, D. (1998) On the Search for the Neural Correlates of Consciousness. In Hameroff, S.R., Kaszniak, A.W., & Scott, A.C. (1998) *Toward a Science of Consciousness II*. 219–230. Massachusetts: MIT Press.

Chalmers, D. (2001) Consciousness and its Place in Nature. Reprinted in Chalmers, D. (eds.) (2010) *The Character of Consciousness*. 103–139. Oxford: Oxford University Press.

Chalmers, D. (2003) How can we Construct a Science of Consciousness? In Gazzaniga, M.S. (2003) *The Cognitive Neurosciences III, second edition*. 1111–1120. Massachusetts: MIT Press.

Chalmers, D. (2010) *The Character of Consciousness*. Oxford: Oxford University Press.

Chalmers, D. (2015) Panpsychism and Panprotopsychism. In Alter, T. & Nagasawa, Y. (eds.) (2015) *Consciousness in the Physical World: Perspectives on Russellian Monism*. 246–276. Oxford: Oxford University Press.

Churchland, Patricia Smith. (1983) Consciousness: The Transmutation of a Concept. *Pacific Philosophical Quarterly*, 64: 80–95.

Churchland, Patricia Smith. (1989) *Neurophilosophy: Towards a Unified Science of the Mind-Brain*. Massachusetts: MIT Press.

Churchland, Paul. (1981) Eliminative Materialism and the Propositional Attitudes. *The Journal of Philosophy*, 78(2): 67–90.

Churchland, Paul. (1990) Knowing Qualia: A Reply to Jackson. In Churchland, P.M., *A Neurocomputational Perspective: The Nature of Mind and the Structure of Science*. 67–76. Massachusetts: MIT Press.

Churchland, Paul. (1992) *A Neurocomputational Perspective: The Nature of Mind and the Structure of Science*. The MIT Press: Massachusetts.

Coleman, S. (2012) Mental Chemistry: Combination for Panpsychists. *Dialectica*, 66: 137–166.

Conduct, M. (2012) Naïve Realism Without Disjunctivism About Experience. *Consciousness and Cognition*, 21: 727–736.

Conee, E. (1994) Phenomenal Knowledge. *Australasian Journal of Philosophy*, 72(2): 136–150.

Cowey, A. (2010) The Blindsight Saga. *Experimental Brain Research*, 200(1): 3–24.

Cox, B. & Forshaw, J. (2012) *The Quantum Universe: Everything That Can Happen Does Happen*. London: Penguin Books.

Crick, F. (1994) *The Astonishing Hypothesis: The Scientific Search for the Soul*. New York: Touchstone.

Damasio, A. (2010) *Self Comes To Mind: Constructing the Conscious Brain*. London: Vintage Books.

Davidson, D. (1970) Mental Events. In Foster, L. & Swanson, J.W. (1970) *Actions and Events*. 207–244. Oxford: Clarendon Press.

Davidson, D. (1988) The Myth of the Subjective. In Davidson, D. (2001) *Subjective, Intersubjective, Objective.* 39–52. Oxford: Oxford University Press.

Davisson, C. & Germer, L.H. (1927) Diffraction of Electrons by a Crystal of Nickel. *The Physical Review*, 30(6): 705–741.

Deiss, S. (2009) Universal Correlates of Consciousness. In Skrbina, D. (2009) *Mind that Abides: Panpsychism in the New Millenium*. 137–158 Philadelphia: John Benjamins Publishing Company.

Dennett, D. (1982) Comments on Rorty. *Synthese*, 53(2): 349–356.

Dennett, D. (1991a) *Consciousness Explained*. London: Back Bay Books.

Dennett, D. (1991b) Two Contrasts: Folk Craft Versus Folk Science and Belief Versus Opinion. In Greenwood, J.D. (ed.) *The Future of Folk Psychology: Intentionality and Cognitive Science.* Cambridge: Cambridge University Press.

Dennett, D. (1995a) *Darwin's Dangerous Idea: Evolution and the Meanings of Life*. London: Penguin Books.

Dennett, D. (1995b) The Unimagined Preposterousness of Zombies: Commentary on T. Moody, O. Flanagan and T. Polger. *Journal of Consciousness Studies 2*(4): 322–326.

Dennett, D. (2003) *Freedom Evolves*. London: Penguin Group.

Dennett, D. (2013) *Intuition Pumps and Other Tools for Thinking*. London: Norton & Company Ltd.

Dennett, D. (2018) Magic, Illusions, and Zombies: An Exchange. [Online] The New York Review of Books. Available at http://www.nybooks.com/daily/2018/04/03/magic-illusions-and-zombies-an-exchange/. [Accessed 10th October 2018]

Descartes, R. (1641) *Meditations on First Philosophy*. Reprinted in Cottingham, J. (1996). Cambridge: Cambridge University Press.

Edelman, G.M. & Tononi, G. (2000) *A Universe of Consciousness*. New York: Basic Books.

Fairhall, S.L., Hamm, J.P. & Kirk, I.J. (2008) Binocular Rivalry Reveals a Dissociation Between the Subjective Experience and Induced Gamma Oscillation. *European Journal of Neuroscience*, 27: 213–216.

Farah, M. (1995) Visual Perception and Visual Awareness after Brain Damage: A Tutorial Overview. In Umultà, C. & Moscovitch, M. (ed.) *Attention and Performance XV: Conscious and Nonconscious Information Processing*. 37-75. Cambridge: MIT Press.

Feigl, H. (1958) *The Mental and the Physical: The Essay and a Postscript*. Minneapolis: University of Minnesota Press.

Feyerabend, P. (1963) Materialism and the Mind-Body Problem. *The Review of Metaphysics*, 17(1): 49–66.

Feyerabend, P. (1975) *Against Method*. In Feyerabend. P. (1993) *Against Method: Third Edition.* London: New Left Books.

Feyerabend, P. (1991) Concluding Unphilosophical Conversation. In Munévar, G. (1991) *Essays on the Philosophy of Paul Feyerabend*. Springer-Science+Business Media, B.V.: Berlin. 487–527.

Feyerabend, P. (1993) “Postscript on Relativism” in Feyerabend, P. (1993) *Against Method: Third Edition.* 268–272. London: New Left Books.

Flanagan, O.J. (1993) *Consciousness Reconsidered*. Massachusetts: MIT Press.

Flanagan, O.J. & Polger, T.W. (1995) Zombies and the Function of Consciousness. *Journal of Consciousness Studies*, 2(4): 313–321.

Fodor, J.A. (1974) Special Sciences (Or: The Disunity of Science as a Working Hypothesis). *Synthese*, 28(2): 97–115.

Foster, J. (2000) *The Nature of Perception*. Oxford: Oxford University Press.

Frankish, K. (2016) Illusionism as a Theory of Consciousness. *Journal of Consciousness Studies*, 23(11–12): 11–39.

Goff, P. (2006) 'Experiences Don't Sum', *Journal of Consciousness Studies*, 13(10–11): 53–61.

Goff, P. (2010) Ghosts and Sparse Properties. *Philosophy and Phenomenological Research*, 81(1): 119–139.

Goff, P. (2017) *Consciousness and Fundamental Reality*. Oxford: Oxford University Press.

Gopnik, A. (2009) Could David Hume Have Known About Buddhism?: Charles François Dolu, The Royal College of La Flèche, and the Global Jesuit Intellectual Network. *Hume Studies*, 35(1–2): 5–28.

Hall, R.J. (2007) Phenomenal Properties as Dummy Properties. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, 135(2): 199–223.

Havlík, A. (2017) Missing Piece of the Puzzle in the Science of Consciousness: Resting State and Endogenous Correlates of Consciousness. *Consciousness and Cognition*, 49: 70–85.

Hinton, J.M. (1973) *Experiences: An Inquiry into some Ambiguities*. Oxford: Oxford University Press.

Howell, R. (2015) The Russelian Monist's Problems with Mental Causation. *The Philosophical Quarterly*, 65(258): 22–39.

Howson, C. & Urbach, P. (2006) *Scientific Reasoning: The Bayesian Approach. Third Edition.* Illinois: Carus Publishing Company.

Hume, D. (1739) *A Treatise of Human Nature*. London: White Hart.

Hume, D. (1748) An Enquiry Concerning Human Understanding. In Hume, D. (1777) *Philosophical Essays Concerning Human Understanding*. London: A.Miller.

Humphrey, N. (2011) *Soul Dust: The Magic of Consciousness*. London: Quercus.

Huxley, T.H. (1874) On the Hypothesis that Animals are Automata, and its History. In Huxley, T.H. (2013) *Science and Culture, and Other Essays*. 199–250. United States: HardPress Publishing.

Jackson, F. (1982) Epiphenomenal Qualia. *Philosophical Quarterly,* 32(April): 127–136.

Jackson, F., Pargetter, R. & Prior, E.W. (1982) Functionalism and Type-Type Identity Theories. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, 42(2): 209–225.

James, W. (1890) *The Principles of Psychology: Volume One*. New York: Henry Holt & Co. Reprinted by New York: Dover Publications (1950).

James, W. (1904a) Does Consciousness Exist? *The Journal of Philosophy, Psychology and Scientific Methods*, 1(18): 477–491.

James, W. (1904b) A World of Pure Experience. *The Journal of Philosophy, Psychology and Scientific Methods*, 1(20): 533–543.

James, W. (1905) How Two Minds Can Know One Thing. *The Journal of Philosophy, Psychology and Scientific Methods*, 2(7): 176–181.

Kant, I. (1781) *The Critique of Pure Reason*. Translated and edited by Guyer. P. and Wood, A.W. (1998). Cambridge: Cambridge University Press.

Kentridge, R., Heywood C. and Weiskrantz L. (1999) Attention Without Awareness in Blindsight. *Proceedings of the Royal Society Biological Sciences*, 266: 1805–1811.

Koch, C. (2012) *Confessions of a Romantic Reductionist*. Massachusetts: MIT Press.

Kripke, S. (1972) *Naming and Necessity*. Oxford: Basil Blackwell Ltd.

Kripke, S. (2011) Vacuous Names and Fictional Entities. In Kripke, S. (2011) *Philosophical Troubles: Collected Papers, Volume 1*. 52–74. Oxford: Oxford University Press.

Kuhn, T. (1962) *The Structure of Scientific Revolutions*. London: The University of Chicago Press.

Ladyman, J., Ross, D., Spurrett, D. & Collier, J. (2007) *Every Thing Must Go*. Oxford: Oxford University Press.

Lakatos, I. (1970) Falsification and the Methodology of Scientific Research Programmes. In (eds.) Worrall, J. & Currie, G. (1978) *The Methodology of Scientific Research Programmes*. 8–101. Cambridge: Cambridge University Press.

Levy, J. (1977) Manifestations and Implications of Shifting Hemi-Inattention in Commissurotomy Patients. *Advances in Neurology*, 18: 83–92.

Libet, B., Wright, Jr, E.W., Feinstein, B. & Pearl, D.K. (1979) Subjective Referral of the Timing for a Conscious Sensory Experience: A Functional Role for the Samatosensory Specific Projection System in Man. *Brain*, 102: 193–224.

Libet, B., Wright, Jr, E.W. & Gleason, C.A (1982) Readiness-Potentials Preceding Unrestricted 'Spontaneous' Vs. Pre-Planned Voluntary Acts. *Electroencephalography and Clinical Neurophysiology*, 54: 322–335.

Libet, B., Gleason, C.A., Wright, Jr, E.W. & Pearl, D.K. (1983) Time of Conscious Intention to Act in Relation to Onset of Cerebral Activity (Readiness-Potential). *Brain*, 106: 623–642.

Libet, B. (2003) Timing of Conscious Experience: Reply to the 2002 Commentaries on Libet's Findings. *Consciousness and Cognition*, 12: 321–331.

Loar, B. (1990) Phenomenal States. *Philosophical Perspectives*, 4: 81–108.

Locke, J. (1689) An Essay Concerning Human Understanding. Reprinted in (ed.) Nidditch, P.H. (1975) *The Clarendon Edition of the Works of John Locke: An Essay Concerning Human Understanding*. Oxford: Oxford University Press.

Lockwood, M. (1989) Mind, Brain and the Quantum: the Compound 'I'. Oxford: Basil Blackwell Ltd.

Lockwood, M. (1993) The Grain Problem. In *Objections to Physicalism*. Oxford: Oxford University Press. 271–291.

Lycan, W.G. (1987) *Consciousness*. Massachusetts: The MIT Press.

Mach, E. (1897) *Beiträge zur Analyse der Empfindungen*. Reprinted in Waterlow, S. (1914) *The Analysis of Sensations and the Relation of the Physical to the Psychical*. London: The Open Court Publishing Company.

Martin, M.G.F. (2002) The Transparency of Experience. *Mind and Language*, 17(4): 376– 425.

Martin, M.G.F. (2004) The Limits of Self-Awareness. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, 120(1): 37–89.

McDowell, J. (1978) On 'The Reality of the Past.' In Hookway, C. & Pettit, P. (1978) *Action and Interpretation: Studies in the Philosophy of the Social Sciences.* Cambridge: Cambridge University Press.

McDowell, J. (2013) Perceptual Experience: Both Relational and Contentful. *European Journal of Philosophy*, 21(1): 144–157.

McGinn, C. (1993) *Problems in Philosophy: The Limits of Enquiry*. Oxford: Blackwell Publishers.

Mele, A.R. (2009) *Effective Intentions: The Power of Conscious Will*. Oxford: Oxford University Press.

Moody, T.C. (1994) Conversations With Zombies. *The Journal of Consciousness Studies*, 1(2): 196–200.

Moore, G.E. (1903) The Refutation of Idealism. *Mind*, 12(48): 433–453.

Nagel, T. (1974) What Is It Like to Be a Bat? *The Philosophical Review*, 83(4): 435–450.

Natarajan, P., Chadayammuri, U., Jauzac, M., Richard, J., Kneib, J-P., Ebeling, H., Jiang, F., Bosch, F-v-d., Limousin, M. Atek, E.J.M., Pillepich, A., Popa, C., Marinacci, F., Hernquist, L., Meneghetti, M. & Vogelsberger, M. (2017) Mapping Substructure in the *HST* Frontier Fields Cluster Lenses and in Cosmological Simulations. *Monthly Notices of the Royal Astronomical Society*, 468(2): 1962–1980.

Neurath, O. (1931) Physicalism: The Philosophy of the Viennese Circle. *The Monist*, 41(4): 618–623.

Overgaard, M. (2011) Visual Experience and Blindsight: A Methodological Review. *Experimental Brain Research*, 209: 473–479.

Parfit, D. (1971) Personal Identity. *The Philosophical Review*, 80(1): 3–27.

Penrose, R. (1989) *The Emperor's New Mind*. London: Penguin Books Ltd.

Pereboom, D. (2015) Consciousness, Physicalism, and Absolutely Intrinsic Properties. In Alter, T. & Nagasawa, Y. (eds.) (2015) *Consciousness in the Physical World: Perspectives on Russellian Monism*. 300–323. Oxford: Oxford University Press.

Pinto, Y., Neville, D.A., Otten, M., Corballis, P.M., Lamme, V.A.F., de Haan, E.H.F., Foschi, N. & Fabri, M. (2017) Split Brain: Divided Perception but Undivided Consciousness. *Brain*, 140: 1231–1237.

Place, U.T. (1956) Is Consciousness a Brain Process? *British Journal of Psychology*, 47: 44–50.

Popper, K. (1935) *Logik der Forschung*. Translated in Popper, K. (1959) *The Logic of Scientific Discovery*. London: Hutchinson & Co.

Popper, K. (1962) *Conjectures and Refutations*. London: Basic Books.

Putnam, H. (1967) The Nature of Mental States. Reprinted in Putnam, H. (1975) *Philosophical Papers Vol 2*. Cambridge: Cambridge University Press.

Quine, V.O. (1951) Two Dogmas of Empiricism. *The Philosophical Review*, 60(1): 20–43.

Quine, V.O. (1969) *Ontological Relativity and Other Essays*. New York: Columbia University Press.

Ramachandran, V.S. (2004) *A Brief Tour of Human Consciousness.* London: Profile Books Ltd.

Ramsay, W. (1990) Where Does the Self-Refutation Objection Take us? *Inquiry*, 33(4): 453–465.

Reppert, V. (1992) Eliminative Materialism, Cognitive Suicide, and Begging the Question. *Metaphilosophy*, 23(4): 378–392.

Rey, G. (1992) Sensational Sentences Switched. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, 63(3): 289–319.

Rey, G. (2016) Taking Consciousness Seriously – as an Illusion. *Journal of Consciousness Studies*, 23(11–12): 197–214.

Rogers, G.A.J. (2004) Locke and Objects of Perception. *Pacific Philosophical Quarterly*, 85: 254–254.

Rorty, R. (1965) Mind-Body Identity, Privacy, and Categories. *The Review of Metaphysics*, 19(1): 24–54.

Rorty, R. (1979) *Philosophy and the Mirror of Nature*. Princeton: Princeton University Press.

Rorty, R. (1982) Contemporary Philosophy of Mind. *Synthese*, 53(2): 323–348.

Rorty, R. (1999) A World Without Substances or Essences. In Rorty, R. (1999) *Philosophy and Social Hope*. London: Penguin Books.

Rorty, R. (2000) Response to Bjørn Ramberg. In Brandom, R. (ed.) *Rorty and his Critics*. Malden, MA: Blackwell.

Rosenberg, G. (2004) *A Place for Consciousness*. Oxford: Oxford University Press.

Russell, B. (1910–1911) Knowledge by Acquaintance and Knowledge by Description. *Proceedings of the Aristotelian Society* 11: 108–128.

Russell, B. (1914) *Our Knowledge of the External World*. London: Routledge Classics.

Russell, B. (1918) On Sensations and Ideas. In Slater, J. (1986) *The Philosophy of Logical Atomism and Other Essays*. London: George Allen & Unwin.

Russell, B. (1921) *The Analysis of Mind*. London: George Allen & Unwin Ltd.

Russell, B. (1927) *The Analysis of Matter*. Nottingham: Spokesman.

Ryle, G. (1949) *The Concept of Mind*. Chicago: The University of Chicago Press.

Satel, S. & Lilienfeld, S.O. (2013) *Brainwashed: The Seductive Appeal of Mindless Neuroscience*. New York: Basic Books.

Sellars, W. (1956) Empiricism and the Philosophy of Mind. In Sellars, W. (1963) *Science, Perception and Reality*. 127–196. London: Routledge & Kegan Paul.

Silvanto, J. (2008) A Re-Evaluation of Blindsight and the Role of Striate Cortex (V1) in Visual Awareness. *Neuropsychologia*, 46(12): 2869–2871.

Skrbina, D. (2009) *Mind That Abides: Panpsychism in the New Millennium.* Philadelphia: John Benjamins.

Smart, J.J.C (1959) Sensations and Brain Processes. *The Philosophical Review*, 68(2): 141–156.

Smith, P. & Jones, O.R. (1986) *The Philosophy of Mind: An Introduction*. Cambridge: Cambridge University Press.

Snowdon, P. (1981) Vision and Causation. *Proceedings of the Aristotelian Society*, 81: 175–192.

Snowdon, P. (2005) Some Reflections on an Argument from Hallucination. *Philosophical Topics*, 33(1): 285–305.

Song, X. & Tang, X. (2008) An Extended Theory of Global Workspace of Consciousness. *Progress in Natural Science*, 18: 789–793.

Speaks, J. (2009) Transparency, Intentionalism, and the Nature of Perceptual Content. *Philosophy and Phenomenological Research*, 79(3): 539–573.

Sperry, R.W. (1968) Hemisphere Deconnection and Unity in Consciousness. *American Psychologist*, 23: 723–33.

Sprigge, T. (1983) *The Vindication of Absolute Idealism*. Edinburgh: Edinburgh University Press.

Stoljar, D. (2005) Physicalism and Phenomenal Concepts. *Mind and Language*, 20(5): 469–494.

Strawson, G. (2006) Realistic Monism – Why Physicalism Entails Panpsychism. *Journal of Consciousness Studies*, 13(10–11): 3–31.

Strawson, G. (2014) *The Secret Connexion: Causation, Realism, and David Hume*. Oxford: Oxford University Press.

Strawson, P.F. (1974) Self, Mind and Body. In Strawson, P.F. (2008) *Freedom and Resentment and Other Essays*. Oxon: Routledge. 186–195.

Stroud, B. (2000) *The Quest for Reality*. Oxford: Oxford University Press.

Swinburne, R.G. (1980) Review of Scientific Realism and the Plasticity of Mind. *Philosophy*, 55(212): 273–275.

Tallis, R. (2011) *Aping Mankind*. Durham: Acumen.

Tartaglia, J. (2007) *Rorty and the Mirror of Nature*. Oxon: Routledge.

Tartaglia, J. (2013) Conceptualizing Physical Consciousness. *Philosophical Psychology*, 26(6): 817–838.

Tartaglia, J. (2016) *Philosophy in a Meaningless Life*. London: Bloomsbury.

Taylor, K.I. & Regard, M. (2003) Language in the Right Cerebral Hemisphere: Contributions from Reading Studies. *Physiology*, 18(6): 257–261.

Thagard, P. & Stewart, T.C. (2014) Two Theories of Consciousness: Semantic Pointer Competition vs. Information Integration. *Consciousness and Cognition*, 30: 73–90.

Tononi, G. (2004) An Information Integration Theory of Consciousness. *BMC Neuroscience,* 5: 42.

Tononi, G. (2005) Consciousness, Information Integration, and the Brain. *Progress in Brain Research*, 150: 109–126.

Tononi, G. & Koch, C. (2008) The Neural Correlates of Consciousness: An Update. *Annual New York Academy of Sciences*, 1124: 239–26.

Trimble, V. (1987) Existence and Nature of Dark Matter in the Universe. *Annual Review of Astronomy and Astrophysics*, 25: 425–472.

Tye, M. (2000) *Consciousness, Color, and Content*. Massachusetts: MIT Press.

Tye, M. (2002) Representationalism and the Transparency of Experience. *Nous*, 36(1): 137–151.

Valberg, J.J. (1992) *The Puzzle of Experience*. Oxford: Oxford University Press.

Velichovsky, B.B. (2017) Consciousness and Working Memory: Current Trends and Research Perspectives. *Consciousness and Cognition*, 55: 35–45.

Von Neumann, J. (1932). *Mathematical Foundations of Quantum Mechanics*. Translated by Bayer, R.T. (1955). Princeton: Princeton University Press.

Walborn, S.P., Terra Cunha, M.O., Pádua, S. & Monken, C.H. (2002) Double-Slit Quantum Eraser. *Physical Review A*, 65: 1–15.

Wartofsky, M.W. (1991) How To Be A Good Realist. In Munévar, H. (ed.) (1991) *Beyond Reason: Essays on the Philosophy of Paul Feyerabend*. 24–40. Dordrecht: Springer Science+Business Media.

Weyl, H. (1949) *Philosophy of Mathematics and Natural Science*. London: Oxford University Press.

Whitehead, A.N. (1925) *Science and the Modern World.* New York: The Free Press.

Wiggins, D. (1967) *Identity and Spatio-Temporal Continuity*. Oxford: Basil Blackwell.

Wilson, J. (2011) Non-Reductive Realization and the Powers-Based Subset Strategy. *The Monist*, 94(1): 121–154.

Wittgenstein, L. (1921) *Logisch-Philosophische Abhandlung*. Translated by Ramsay, P. (1922) in *Tractatus Logico-Philosophicus*. London: Kegan Paul.

Wittgenstein, L. (1953) *Philosophical Investigations*. Translated by Anscombe, G.E.M (1958). Oxford: Basil Blackwell.

Wooters, K.W. & Zurek, W.H. (1979) Complementarity in the Double-Slit Experiment: Quantum Nonseparability and a Quantitative Statement of Bohr's Principle. *The Physical Review*, 19(2): 473–484.

Young, T. (1804) The Bakerian Lecture: Experiments and Calculations Relative to Physical Optics. *Philosophical Transactions of the Royal Society of London*, 94: 1–16.