Consciousness, Representation, and Flexibility

by

William Languedoc

A thesis submitted to the Faculty of Graduate and Postdoctoral Affairs
in partial fulfillment of the requirements for the degree of

Master of Arts

in

Philosophy

Carleton University
Ottawa, Ontario

Abstract

Determining whether another being is conscious first involves determining what consciousness *is*. In Chapter 1, I argue against the view that consciousness is unlike the rest of cognition, and in favor of a view that sees consciousness as a cognitive function like any other. In Chapter 2, I argue that most cognitivist accounts of consciousness fail in one of two ways: they are narrowly circular, or they change the subject. I suggest that one way to avoid these common routes to failure is to adopt what I refer to as the Minimal Claim – that consciousness has a representational base. In Chapter 3, I apply the work done in the two previous chapters to the issue of ascribing consciousness to beings other than ourselves, suggesting that behavioral flexibility may play an important role.

Acknowledgments

I would like to extend my deepest thanks to all those who made a contribution, directly or indirectly, to the completion of this thesis:

To my supervisor, Professor Andrew Brook, for his expert guidance, thoughtful critiques, and commendable patience throughout the development of this work.

To my family – Geoffrey, Margot, Jen, and my partner Ellen. This endeavor would not have been possible without all of your loving support.

To Professors Myrto Mylopoulos, Christine Koggel, David Matheson, and Eros Corazza, for valuable discussions and insights along the way.

And to Carleton University, the Carleton University Philosophy Department, and the David and Rachel Epstein Foundation, for their generous financial support.

Table of Contents

Introduction

I have two primary goals in putting together this thesis. The first is to establish what the term 'consciousness' actually ought to be used to refer to. This is because for a term that is used so often both in academic and conversational settings, there is considerable lack of agreement on what "consciousness" ought to apply to. In most non-academic settings, to be "conscious *of*" something is to be aware of it. In a sentence like "we ought to be conscious of the effect that our daily habits have on climate change", what is typically meant is that we should acknowledge, or be aware of, said habits. Looking more closely, a conscious thought in this case may also be connotated with a kind of consistent awareness – keeping the thought at the "front of one's mind", so to speak. In cases like these, I suspect "conscious" is simply used as a slightly-more-sophisticated-sounding word for "aware". In the same sense, if someone has a "sub-conscious" thought or desire, we understand that to mean that the thought or desire is present, and perhaps playing a role in their decision-making process, but unacknowledged by or unknown to the person to whom it belongs. Another common use of the term is in evaluating responsiveness and lucidity. If you come across someone who is lying unresponsive on the ground, you might say to the arriving paramedics that the person was "unconscious" when you found them. A statement like this would be understood as meaning that the person was unresponsive and immobile. Similarly, when someone is fast asleep, or under anaesthesia, you might say that they are in an unconscious state, due to their lack of responsiveness to external stimuli. I have also seen the term used rather frequently in the media, in the form of "social/cultural/global/collective consciousness". In this case it seems that what is being

referred to is simply shared knowledge. When a reporter claims that terrorism has made its way into the cultural consciousness of the USA since the events of 9/11, for instance, what they mean is that since the events of 9/11, the USA has, as a culture, come to be aware of the dangers and probability of terrorism. It seems that in all common uses of the term, "consciousness" is intended to relate in some way to awareness, whether it be to external stimuli or to the forces that guide decision making processes.

In academic settings, consciousness tends to take on a more complex, and certainly more hotly contested, meaning. It is such a controversial term that any attempt to explain how consciousness works necessitates a preamble that clearly outlines, and argues for, the definition of consciousness that the author is using. Further complicating the matter is the truly vast number of "x-consciousness" -style terms that have been generated by philosophers and scientists who have tried to make sense of the issue:

> access consciousness, phenomenal consciousness (Block 1995), psychological consciousness (Chalmers 1996), self-consciousness, first-order (aka simple) consciousness (Dretske 1995), creature consciousness (both transitive and intransitive), state consciousness (both transitive and intransitive) (Rosenthal 1992), monitoring consciousness, reflective consciousness, prereflective consciousness, background consciousness, focal consciousness, peripheral consciousness, fringe consciousness (various including Gallagher & Zahavi 2007), conscious awareness (and what is unconscious awareness supposed to be like?) (Chalmers 1996), qualia (many writers), unified consciousness, macroconsciousness, microconsciousness Zeki 2003), afferent, perceptual, and reflective consciousness (Zlatev 2008, Honderich 2006),

> observational, introspective and intuitive consciousness (Itkonen 2008), not to mention Tye's (2003) D-, I-, P-, and R-consciousness. (Brook 2019)

This constantly growing list of terms is reflective of a considerable lack of agreement on what consciousness is, what it is constituted of, and how it is realized. The controversial nature of the concept of consciousness justifies an approach to defining it that starts at the most basic level, which can then be progressively built upon. As a result, I have decided to adopt a 3-part structure for my investigation into the nature of consciousness. In the first two chapters, I will lay out the general kind of thing that I take consciousness to be (a cognitive phenomenon), and elaborate on the kinds of cognitive mechanisms that I take to underly it. In the third chapter, I will connect my definition of consciousness to its most important neighbour – externally observable behavior.

I will take Chapter 1 to focus on determining whether there is anything to the many claims that consciousness is a non-cognitive phenomenon. Despite living in a time when dualism is an unpopular position to take in academia, and a charge of epiphenomenalism verges on being an insult, there are numerous philosophers who argue that consciousness cannot be explained through an appeal to cognition, and who instead propose theories that rest upon immaterial substances and properties. I will evaluate some of the most influential arguments for the non-cognitive nature of consciousness, and provide a series of critiques that I hope will cast enough doubt on these arguments to make them seem highly implausible in comparison to their cognitivist alternatives.

In Chapter 2, I will say more about what I take consciousness to be constituted of. I will argue for the claim that consciousness has a *representational base*. The notion of the representational base of consciousness was originally developed by Andrew Brook, and it

is an idea that I believe is likely to find widespread agreement. In saying that consciousness has a representational base, what I mean to assert is that mental representation is a necessary condition for consciousness. I will argue that this claim, which I refer to as the Minimal Claim, is the most that we can say about what constitutes consciousness, at least given the current state of the art.

In Chapter 3, I will explore the question of how we might leverage the notion of the representational base to investigate non-verbal (and in this case, non-human animal) behavior, and its relationship with consciousness. The main issue with placing emphasis on mental representation in evaluating consciousness is that it is notoriously difficult to determine whether some behavior is facilitated by conscious, rather than unconscious, representation. I propose that behavioral flexibility is a characteristic of behavior that offers some promising footholds that we can use to identify conscious representation. I also propose that, rather than looking for behavioral indicators of consciousness *per se*, we ought to conceptualize such a process as looking for behavioral indicators that establish a probability of consciousness. This soft approach to investigating indicators of consciousness is perhaps less bold than others, but I aim to show that it is the most that we can expect to accomplish given the epistemic issues and scientific limitations of our time, and that it nonetheless provides a very useful starting point.

Chapter 1: The Ontology of Consciousness

Consciousness has recently received quite a bit of interest in the philosophical community, particularly since the mid 1990's, when a number of important philosophical works were published on the subject. Chalmers' entry in the Journal of Consciousness Studies, "Facing up to the Problem of Consciousness", articulated the easy and hard problems of consciousness, which are just as relevant today as they were 25 years ago. Dennett's "Consciousness Explained", which was published in 1991, effectively tackled some long-held beliefs on the topic of consciousness and epistemology and introduced a deep running breed of skepticism to the field that was largely missing up until that point. Since then, there has been a massive production of scholarship on the subject. Despite the increased interest, progress has been slow, and determining what consciousness actually is, and how it happens, continues to be the dominant question addressed in the literature. In this chapter, I will give my own evaluation of some of these efforts, focusing on the theories that to my mind have had the greatest impact on the field.

1.1 – The Case for Anti-Cognitivism

One of the most foundational disagreements amongst philosophers of consciousness has to do with the question of whether a full account of consciousness can be provided by science. This is not a dispute as to whether the science of *today* can explain consciousness – I don't think that anyone would try to claim that science has already, or even will very soon, give us the answers that we're looking for. This is more of a theoretical

dispute relating to the definition of science itself. Science is, typically, concerned with objective facts. When someone undertakes the scientific study of colour, for instance, they look at data related to wavelength. Similarly, when looking at pain, a researcher will examine the way that information is transmitted through nerve endings and into the brain. What is not studied, and according to some is beyond the scope of science, is *experience*. The experience of colour (how it appears to someone) and the experience of pain (how pain actually feels) are phenomenal, and according to some cannot be fully accounted for by simply explaining the corresponding brain states and wavelengths.

On one side of the dispute, there are the cognitivists, who hold that "consciousness is similar to other cognitive functions, such as representation, memory, and reasoning" (Brook & Raymont). Cognitivists see consciousness as being describable in purely cognitive terms. If memory can be explained scientifically and objectively, without leaving out any important features, then in theory consciousness should be similarly explainable. The result of treating consciousness as a kind of cognition is that a completed neuroscience, or cognitive science, should be able to give a purely scientific account of consciousness without leaving its phenomenal qualities unaccounted for.

On the other side, there are anti-cognitivists, who hold that consciousness is completely dissimilar to any cognitive functions. Anti-cognitivists argue that if anything might eventually provide an explanation for the phenomenal quality of consciousness, it will involve an explanation that goes beyond cognitive processes, and that it is perfectly conceivable for complex cognition to take place in the absence of any kind of consciousness.

In the following, I will explain the approaches of some of the most prominent contemporary anti-cognitivists (Nagel, Block, and Chalmers). Afterwards, I will provide an analysis and critique of the anti-cognitivist position in general, arguing in favor of a

cognitivist perspective on the grounds that anti-cognitivists have not provided reasons to adopt their theories that meet the high threshold set by Occam's Razor.

Thomas Nagel – What It Is Like-ness

Nagel's seminal paper *What Is It Like to Be a Bat?* (1974) laid much of the groundwork for discussing phenomenal consciousness. Nagel begins his paper with a bleak analysis of the reductionist study of consciousness at the time. In the 1970's, and really up until the early 1990's, consciousness was either completely ignored as an intractable problem, or dealt with through reductive analogy, with questionable results. On the state of consciousness research, Nagel writes, "philosophers share the general human weakness for explanations of what is incomprehensible in terms suited for what is familiar and understood, though entirely different." (Nagel 1974 P.435).

Nagel acknowledges that consciousness of one kind or another is widespread and likely comes in "countless forms totally unimaginable to us" (P.436). On this point I certainly agree with him (though unlike Nagel I think that these forms are invariably cognitive), and developing an account of what these "forms" of consciousness might consist of is the primary goal of my project. The way that Nagel justifies this belief in the widespread existence of consciousness is through a tantalizingly simple definition of what the subjective character of experience *is*. Nagel writes,

> But fundamentally an organism has conscious mental states if and only if there is something that it is like to *be* that organism – something it is like *for* the organism. (p.436)

Nagel's definition highlights the problem with many attempts to define consciousness through reduction to brain states. It seems that many reductive accounts can be fully fleshed out without ever having to deal with experience, phenomenology. But the issue with that is that if we are not discussing experience, we seem to have changed the subject from consciousness to something else (for a more detailed discussion of subject changing, see Chapter 3). For example, one of the reductive analogies that Nagel references at the beginning of his paper is the "Water/H_2O" reduction. In this analogy, water is to consciousness as H_2O is to brain states. If water is reducible to H_2O, then the analogy goes that consciousness should therefore be reducible to brain states. But again, to use an analogy of this sort makes it easy to change the subject, and thereby lose exactly the feature of consciousness that is in need of explanation – its phenomenal character.

An important sticking point regarding Nagel's approach is how to interpret "something it is like to *be*". On one interpretation, for there to be something it is like to be an organism, self-consciousness is required. This would mean that consciousness of the external world is possible, but reliant upon consciousness of self. In order for me to be conscious of a bug flying across my field of view, I would have to also be conscious of myself as the experiencer of this vision. This would seem to disqualify most, maybe all, non-human animals from having consciousness, since most accounts of consciousness of self involve some significant conceptual framework. The alternate interpretation of Nagel's characterization of consciousness is that "'it being like something', describes "*a way in*

which something is presented to someone." (Brook and Raymont, p.38). On this interpretation, we can say that there is something it is like for me to be conscious of a bug flying across my field of view, without my being conscious of myself as being in some mental state. I take this to be the more plausible of the two interpretations – there does not appear to be any *prima facie* reason why consciousness of the self would be required in order to have experience of the external world, or even internal states such as pain or hunger. I take the latter interpretation to be the stronger and more inclusive of the two, because it is the only one that allows for non-human animal consciousness.

Ned Block – Phenomenal and Access Consciousness

Block argues that much of the confusion surrounding consciousness stems from a conflation of two very different concepts. He begins his paper *Concepts of Consciousness* (Block 2002) by comparing this conflation with the way that Aristotle used the term "velocity" to refer to both average and instantaneous velocity – two very different concepts that when treated as one, have the result of generating paradoxical and confusing results. In the case of consciousness, Block believes that the confusion comes from a failure to distinguish between *access consciousness* (A-consciousness), and *phenomenal consciousness* (P-consciousness).

P-conscious states are those that have the quality of there being "something it is like" to be in said state. Block acknowledges that the problem with a definition of this sort, and indeed there is the same problem with Nagel's definition of "what it is like-ness", is its circularity. In the case of P-consciousness, words such as "experience", "experiential

properties", and "experiential states", are all used as terms to define what makes a state P-conscious. The circularity comes from the fact that all of these terms are close to being synonymous with "phenomenal", and therefore tell us nothing that we did not already know. Block's description of P-conscious states is decidedly anti-cognitive, as he argues that the properties that make a state P-conscious are "distinct from any cognitive, intentional, or functional property" (Block 1995, p.230).

A-conscious states are those that have the characteristic of being "inferentially promiscuous"; that is, broadcast and made available for use as a premise in reasoning and poised for rational control of action and speech. Block leaves significant room for flexibility in this definition – for instance he does not believe that A-conscious states must necessarily be connected to speech in any way, since he wants to allow for the possibility of non-linguistic animals having them (Block 2002, p.208). An important feature of A-consciousness is that it is representation-based. It involves a mental representation of *x* being broadcast for use in reasoning. A-conscious states are also system dependent. A state becomes P-conscious as a result of "what happens *inside* the P-conscious module" (Block 1995, p.232). A state becomes A-conscious not in virtue of what happens inside of a particular module, but in virtue of the relations between modules, since it is this relationship that allows for the Executive System involved in rational decision making to have access to the content of the representation. In short, Block thinks that the distinction between the two kinds of consciousness is so deep-running that P-consciousness is a product of its own module, and A-consciousness is a mode of relationship and causal access among various other elements of cognition.

Block claims that a property of P-consciousness is that it is subject to the "explanatory gap". Quoting T. H. Huxley (1866), he writes,

> How it is that anything so remarkable as a state of consciousness comes about as a result of irritating nervous tissue, is just as unaccountable as the appearance of Djin when Aladdin rubbed his lamp. (p.210)

Here Huxley is articulating a problem that remains unresolved, and has been a thorn in the side of consciousness theorists ever since: how can consciousness arise from things that are not themselves conscious (neurons, for instance)? Much of Block's writing is firmly grounded in talk of neuroscience and "material" accounts of consciousness, though there are claims that he makes throughout his writing that hint at, and sometimes make explicit, his skepticism about cognitive definitions of consciousness. Alluding to Chalmers' zombie thought experiments, he asks "[w]hy couldn't there be brains functionally or physiologically just like ours…whose owners' experience was different from ours or who had no experience at all?" (Block 1995, p.231). In the next section, I will provide an argument as to why I believe this to be a conceptual impossibility.

David Chalmers – Problems of Consciousness and Zombies

David Chalmers is perhaps most well-known for formulating the "easy and hard problems" – two problems that he claims must be solved in order to unravel the mystery of consciousness. The "easy problem" is the question of how to explain the causal role that certain cognitive functions play in the production of behavior (Chalmers 2003, p.103). It is easy in the sense that in order to explain this causal relationship, all that needs to happen is for a particular mechanism within the brain to be identified as the

cause of *x* behavior. According to Chalmers, there is good reason to believe that cognitive functions can play these causal roles, and determining exactly what these relationships consist of is a conceptually straight-forward problem. Behavior is observable and quite clearly physical, and so are brain states, so linking the two should require no metaphysical gymnastics.

The "hard problem" is presented as being more intractable. In order to solve it, what must be explained is how and why some physical process in the brain results in experience, or the feeling of there being "something it is like". Chalmers writes,

> [e]ven once one has an explanation of all the relevant functions in the vicinity of consciousness – discrimination, integration, access, report, control – there may still remain a further question: why is the performance of these functions accompanied by experience? (p.104)

Unlike the observable behavior that must be explained in the easy problem, the phenomenal nature of consciousness makes it difficult to study, and also sets it apart from the rest of the cognitive functions of the brain that do not have the characteristic of there being "something it is like" to be in *x* state. On Chalmers' view, when we try to connect conscious experience to cognitive function, we run into exactly the same mind-body and mind-mind problems that dualists have run into for centuries in explaining how the immaterial soul could have any causal relationship with a material body, and vice versa.

It is important to note that, although Chalmers was the first to use the term "hard problem", he was not the first to articulate the question of how consciousness could arise from non-conscious material. In his quote comparing consciousness arising from nervous

tissue to the Djin from Aladdin's lamp, Huxley was quite explicitly raising the question that Chalmers' hard problem would later be based on. Additionally, almost two centuries before Huxley was even born, Gottfried Leibniz made a very similar point in his *Monadology:*

> [S]upposing that there were a mechanism so constructed as to think, feel and have perception, we might enter it as into a mill. And this granted, we should only find on visiting it, pieces which push one against another, but never anything by which to explain a perception. (Leibniz 1714 – Remark 17)

One way to illustrate this well-established problem is through the example of a phase shift. For instance, we know a lot about the way that H2O shifts from being a liquid to a gas. Scientists have been able to tell this story in great detail – when H2O is subjected to certain temperature conditions the molecular bonds will weaken, causing a phase change from liquid to gas. We cannot tell the same kind of story about how neuronal activity in the brain creates consciousness. Additionally, we can say that it is required that the phase change happens under certain circumstances, since when water is heated to above a certain temperature, the phase shift is inevitable. We do not have any kind of similar understanding of what the conditions are under which consciousness necessarily arises. Even more fundamentally, though, we understand what liquids are, and what gasses are, and what it looks like when a phase change happens. We don't have any idea what it looks like for neurons to move from being non-conscious to conscious (or whether that's even the right kind of question to ask), because we don't understand what the basic

qualities of consciousness are. We can't begin to really understand the conditions that make consciousness possible until we understand what it is exactly that consciousness *is*.

Chalmers illustrates this problem with the "conceivability argument", more commonly known as the "zombie" thought experiment (Chalmers 2002, p.249). The argument postulates the existence of a system that from a third-person and first-person perspective, resembles that of a conscious being in every way, and yet lacks consciousness. You would be able to interact with it in all of the same ways that you might interact with another conscious being, and in all cases its behavior would be indistinguishable from another conscious being. If this system took a human form, its brain states would even be exactly identical to the brain states of a conscious human's brain, however it would still lack consciousness. The argument goes that if zombies are conceivable, then they are at least metaphysically possible. If they are metaphysically possible, then consciousness cannot be exhaustively defined through an appeal to cognition.[1] Chalmers takes cognitivist approaches to consciousness as being not just incorrect, but *incoherent* – largely because of the conclusions that he believes his zombie thought experiment results in. In his argument against cognitivism, Chalmers provides three anti-cognitivist alternatives that he believes have the upper hand by being, at minimum, coherent.

The first is *interactionism*. On this view, cognitive states and phenomenal (mental) states are ontologically distinct, and yet can interact and have a causal relationship. The obvious objection to this view is that there is no plausible "causal nexus" between the physical and phenomenal – one cannot locate a particular part or function of the brain

[1] There has been a considerable amount of literature produced on the topic of conceivability, none of which I will explore in this chapter. It is worth noting, however, that it is not universally agreed upon that conceivability results in metaphysical possibility, or that metaphysical possibility can tell us anything meaningful about the nature of things (Dennett 1991, p.401).

that might allow for any kind of communication between the two different kinds of states. Not that this really matters anyway, however, since traditional physics tells us that such a causal relationship between the physical and the non-physical would be impossible. Chalmers claims that the strongest response to this objection, and one that he seems to personally support, is that modern quantum physics is in fact quite friendly to an interactionist picture of consciousness (p.262). The basic idea behind this response is that there is an uncertain amount of physical nondeterminism at play at the subatomic level, whereby the act of observation has been shown to cause a "collapse" of the superposed wave function used to describe the state of matter (Schrödinger Dynamics). To Chalmers, this observation-caused collapse suggests that the non-cognitive could play a causal role.

The second alternative view is *epiphenomenalism*. On this view, phenomenal and cognitive properties are ontologically distinct, but phenomenal properties can have no causal impact on cognitive properties. cognitive states, however, do give rise to coinciding phenomenal states. Chalmers does not give too much credence to this view, as he takes it to be inelegant and counter intuitive. Explaining why we have experience - why cognitive states give rise to phenomenal states – is no easier to explain on the epiphenomenal view than it is on the cognitivist view. Chalmers chooses to include it in his three alternatives because he takes it to not have any fatal, knock-down arguments against it. Taking a jab at cognitivism, he writes "inelegance and counterintuitiveness are better than incoherence" (p.265).

The third and final alternative view is *panprotopsychism*, which is the view that "consciousness is constituted by the intrinsic properties of fundamental physical entities" (Chalmers 2002). Panprotopsychism is importantly distinct from pansychism; pansychists posit that all matter has phenomenal properties, whereas panprotopsychists posit that all

matter has special proto-phenomenal properties, which on their own do not have any kind of phenomenal character, but in the right combinations give rise to phenomenal experience (Chalmers 2013). It is distinct from cognitivism in that these proto-phenomenal properties are themselves non-cognitive, despite being properties of the cognitive system. This can be thought of as a kind of melding together of anti-cognitivism and cognitivism.

1.2 – The Case Against Anti-Cognitivism

The Failure of Zombie Thought Experiments

Zombie thought experiments are often used as arguments against the plausibility of consciousness being a strictly cognitive phenomenon. They are certainly intuitively appealing – we have no trouble imagining a computer as being highly intelligent and functionally similar to a conscious being, despite lacking consciousness altogether. This makes conceptualizing a "zombie" that has similar characteristics not too much of a stretch. In the thought experiment, one is made to imagine a molecule-for-molecule clone of a conscious being's brain. The only difference between the two brains is that the original has consciousness, while the clone does not. From here it is argued that if conceptualizing such a circumstance is possible, then it must be the case that the existence of consciousness in some system is not dependent only on the cognitive makeup of that system. It may be the case that such a system is not practically possible – one could argue that there is no practical way to make a molecule-for-molecule clone of a conscious creatures brain and not wind up with a conscious being. However, proponents of this thought experiment will

be quick to argue that what is needed is not natural possibility, but *metaphysical* possibility. They argue that we can imagine, however implausibly, that a zombie is possible, but that we could not in the same way imagine a square circle, for instance, because a square circle is a metaphysical impossibility, while zombies are not.

I believe that these thought experiments are so appealing because they take advantage of some powerful intuitions that many of us have. Many people think that consciousness is tied to phenomenal states that may be taken to be non-cognitive. It is what makes us "who we are", so to speak, and the idea that this could be a straightforwardly cognitive phenomenon is unsettling for some. So when asked to conceptualize a perfect clone of your brain, those who already have this bias towards consciousness as a non-cognitive phenomenon will have no trouble conceiving of a being in a state where "the lights are on but there's no one home". The trouble is that this kind of argument is only convincing, or even intelligible, to those of us who already are under the impression that consciousness is non-cognitive. If you believe, for whatever reason, that consciousness is a purely cognitive phenomenon, then the idea of a perfect clone of your brain that nonetheless lacks some essential physical property of your cognitive ability (namely consciousness) is completely self-contradictory. I do not take this to be an argument for cognitivism, necessarily, but instead that this thought experiment cannot function well as an argument for either side of the debate.

Dennett also has much to say about zombie thought experiments, and uses Rosenthol's HOT theory to demonstrate their implausibility. He proposes the idea of a "zimbo", which he describes as "a more realistic and complex zombie, which monitors its own activities, in an indefinite upward spiral of reflexivity" (Dennett 1991, p.310). Unlike the paradigmatic zombie of traditional thought experiments, zimboes would have

unconscious higher order thoughts, allowing them to introspect and reflect upon past experiences and decisions. Dennett claims that zimboes should not be an issue for those who believe that the idea of philosophical zombies is coherent, since such an entity would accurately represent the concept of a supposedly non-conscious system that is cognitively identical to that of a conscious being. As such, it would certainly think that it was conscious, despite not having any higher-order thoughts. In Dennett's words, a zimbo "would be the "victim" of the benign user illusion of its own virtual machine", or in other words, its belief in itself as a conscious being would be the result of illusory non-conscious processes (P.309 - 314). The issue here is that on Rosenthal's account, a higher-order thought need not itself be conscious in order to generate a conscious mental state, so by definition the zimbo would in fact be conscious. Once again, the idea of a philosophical zombie becomes metaphysically impossible.

Dennett's analysis specifically targets Rosenthal's approach, but the critique that he provides is actually quite general. Zombies, if understood as resembling a conscious being in every way from a *third person* perspective only, don't tell us anything meaningful about consciousness. If we are going to discuss zombies honestly, we need to include the zombie's first-person perspective, the zombies' beliefs about itself, the zombies' expressions of feelings, and so on, but at that point there seems to be no difference between the zombie and a conscious creature. A zimbo would act like a conscious creature, as well as think that she is conscious in the same way that we do. On Dennett's corrected definition, we could all be zombies, and at that point the thought experiment cannot be done (p.429).

I believe that a similar issue may be present in Frank Jackson's "Mary the Color Scientist" thought experiment. Jackson illustrates this thought experiment in the following way:

> 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. She discovers, for example, just which wave-length combinations from the sky stimulate the retina, and exactly how this produces *via* the central nervous system the contraction of the vocal chords and expulsion of air from the lungs that results in the uttering of the sentence 'The sky is blue'...
>
> 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, p.130)

This I believe to be one of the most glaring examples of flawed intuitions at work.[2] Jackson claims "it seems just *obvious* that she will learn something new…". But is this really so

[2] To be fair, I should note that Jackson himself no longer believes that this thought experiment is successful (2007). Despite Jackson's change of mind, however, it is still a commonly used and referenced thought experiment, and as such still warrants discussion.

obvious? Consider that what has been said so far is that Mary possesses *all* of the physical knowledge related to colour. She doesn't just have knowledge gained from books and lectures and academic papers, she supposedly has all of the physical knowledge that might ever be associated with an understanding of colour. If this is the case, why will she learn anything new when she steps out the door, or watches colour television? Jackson thinks that it is obvious that she will because he thinks that there is something more than physical about the experience of colour. If you believe that experience can be explained in purely cognitive or physical terms, then you will not find Jackson's conclusion to be quite as obvious as he does, or even coherent.

Thought experiments are useful at discovering our biases and intuitions, but it is not necessarily the case that truths about consciousness must line up with our intuitive responses, so their ability to function as proofs is rather limited. Whether I find the workings of an internal combustion engine to be intuitive or not is of absolutely no consequence to how an internal combustion engine actually works, and the same is true of consciousness. In the words of Dennett, Jackson's example (and I believe Chalmers' too) is "a classic provoker of Philosopher's Syndrome: mistaking a failure of imagination for an insight into necessity" (Dennett 1991, p.401).

Ontological Simplicity

Occam's Razor is the term for the usually correct idea that one's ontology should include as few kinds of entities as possible, because it should be restricted to kinds of entities that we have reason to believe in. In the case of consciousness, and the ontological distinction between the mental and the physical, Occam's Razor would tell us that if

there is any way that we can explain consciousness without introducing non-cognitive substances or properties, then this would be the correct solution. On the topic of ontological simplicity in the context of this debate, Brook and Raymont write:

> Both camps agree about most claims about cognition. The anti-cognitivists argue that consciousness is something more. So they have to make the case. Absent some reason to believe that consciousness is something more than cognition, the rational thing to believe would be that it is cognition. If we can believe in one thing for which we have no evidence, there is nothing to stop us from believing in ten – or ten thousand. This is just Occam's Razor. (Brook and Raymont)

Brook and Raymont raise an important point about multiplying entities. One of the reasons why Occam's Razor ought to be taken so seriously is that it is holding back a potential avalanche of distinct substances and properties. In conversation, Brook has used the analogy of leprechauns dancing on his desk. Just because I cannot prove with certainty that there are not leprechauns dancing on Brook's desk does not mean that I should believe that there are. There is still a large amount of evidence that suggests that there are no dancing leprechauns (no visual evidence, no auditory evidence, they could not be detected by any existing technology, etc.), in the same way that there is a large amount of evidence that suggests that there is no mysterious non-cognitive process occurring in the brain. If, due to the lack of complete certainty with which I can claim that there are no leprechauns dancing on Brook's desk, I choose to believe that there are in fact leprechauns dancing on Brook's desk, what is to stop me from believing in hundreds or thousands of other entities or substances for which there is no evidence? The answer is *nothing*. Occam's Razor serves as a powerful rebuttal to anti-cognitivist arguments

because it highlights the fact that an anti-cognitivist that proposes 1,323,476 different non-cognitive properties of the brain is proposing something no more ridiculous or absurd than someone like Chalmers, who might only be proposing one.

Since we have so far been able to formulate cognitive explanations for many other functions of the brain, I believe there is good reason to think that sticking with a cognitive approach should be the default position. However, if there was an anti-cognitive approach that seemed far more plausible that the cognitive theories that are available today, and provided some sort of huge advantage in terms of explanatory usefulness and accuracy, then we may have good reason expand our ontology. In this section, I will examine Chalmers' three alternatives to cognitive definitions of consciousness to determine whether any of them can meet this threshold of plausibility.

First there is interactionism. This Descartes-inspired approach proposes two distinct substances that have a two-way causal relationship, such that the mental can communicate with the physical, and the physical with the mental. The way that Chalmers proposes that this might happen is through the causal relationship between observation and wave collapse at the quantum level. I don't think that this kind of explanation is particularly attractive. Quantum physics is a field that we still have much to learn about, and as such there is a great deal of disagreement and difference in interpretation amongst physicists about what is actually going on in these experiments. That being said, this is not necessarily an anti-cognitive position, since potential answers from quantum theory could wind up being perfectly compatible with traditionally cognitive explanations of consciousness. Until quantum physicists reach some level of agreement about how to interpret and apply their findings, however, I do not think that we should be using it to justify any particular viewpoint in philosophy.

The second alternative Chalmers gives us is epiphenomenalism. He does not spend a great deal of time trying to justify this view, since he is mainly proposing it because he does not see any "fatal flaws" in it, unlike his view of cognitivism. While this may be true, I do not think that it offers any kind of improvement in plausibility or coherence over a purely cognitive explanation. For one thing, it seems to eliminate the possibility of conscious decision making. If mental states have no causal power, then they are useless at guiding behavior. When I think to myself "I want to get a haircut", and subsequently book an appointment with my barber, there seems to be evidence of conscious decision making. Is this just an illusion? Perhaps - but if it is, the burden of proof lies on the epiphenomenalist to provide compelling reasons to think so.

There are two other reasons why I take epiphenomenalism to be a non-starter. The first relates to the fact that, as far as we know, the physical world is "causally closed". That is, in order for an object to have any kind of physical effect on another physical thing, the object must itself be physical (the issue of causal closure will be dealt with in greater detail later in this chapter, in my discussion of the mind-mind and mind-body problems). Therefore, if the "mental" is not straightforwardly cognitive, then there is no room for it to have any effect on the brain, and vice versa. This counter-argument of course could be applied to any ontology that includes non-physical entities. Unless one intends to propose some process through which the gap between physical and non-physical could be bridged (typically a god would be invoked here), epiphenomenalism has no leg to stand on.

But perhaps the most definitive "nail in the coffin" for epiphenomenalism is the idea that if epiphenomenalism were true, we would have absolutely no way of knowing it. If something is epiphenomenal, by definition that means it is not causally affecting

anything else, including our perceptions and our beliefs. The epiphenomenal state would be unknowable because it wouldn't have effects on us, including the kinds of effects that are required for knowledge. This makes any formulation of epiphenomenalism both unprovable and unfalsifiable, and as a result, I do not think that they are worth taking seriously. Chalmers' half-hearted defense of this view relies on the claim that epiphenomenalism is coherent, while cognitivism is not. Based on the evidence that has been provided in favor of epiphenomenalism (none), and the many good arguments against it, I find it to be a decidedly implaisible view.

The third alternative is panprotopsychism. This view seems to solve the hard problem completely, since it postulates that phenomenal consciousness is just a special, non-cognitive property of matter when it is arranged in the correct ways. It also could provide a somewhat surprising answer to Chalmers' zombie thought experiment. If all matter has proto-phenomenal properties (the potential to be a part of a system that generates conscious experience), and the physical makeup of the zombies brain is such that it is a molecule-for-molecule clone of a conscious being's brain, then it should be impossible for phenomenal consciousness *not* to be generated in the zombie's brain, assuming that the physical construction of some system is what determines whether or not proto-phenomenal properties manifest into phenomenal properties.

I take this alternative to be by far the most likely of the three, and indeed, so does Chalmers. Panprotopsychists and most cognitivists share the belief that consciousness is a product of the entire system, but disagree on whether this process is the result of a material (cognitive) property, or an immaterial property. The views are surprisingly similar, but the cognitive approach once again has the upper hand of ontological simplicity. Panprotopsychism also shares many problems with every other approach to

consciousness. For instance, none can provide a plausible account of how and why particular kinds of systems facilitate phenomenal experience, while others don't. Chalmers acknowledges that nobody really knows exactly how panprotopsychism might work, or what protophenomenal properties actually are, and argues that "our ignorance about protophenomenal properties should not be mistaken for an objection to the truth of panprotopsychism" (Chalmers 2013, p.17). I agree with Chalmers, but surely, the same could be said in favor of cognitivism! If we are at peace with the fact that whatever theory we might propose about the workings of consciousness will have a degree of uncertainty to it, such that we must wait for a "complete" version of said theory to provide the answers, what possible reason might there be to accept a theory that postulates the existence of non-cognitive substances over a theory that does not, if they both provide similarly uncertain answers?

Occam's Razor tends to work best when comparing two theories that agree on many of the essentials and disagree on one element, where one party says "*kind x exists*" and another party says "*kind x doesn't exist*". In cases like these, Occam's Razor says that the onus is on the party postulating the extra element to provide proof of why the element is necessary. If I, taking the stance of an atheist, agree with a religious person on everything but the forces behind existence (i.e. God), then the onus is on the religious person to provide an argument for their position. In other words, the default position would be the atheistic one that I would be inclined to take, and I would be under no obligation to disprove the beliefs of my interlocutor. Applying this notion to the debate at hand, Occam's Razor is particularly effective against panprotopsychism because of the large amount of agreement between cognitivists and panprotopsychists. The only place where the two theories diverge substantially is in their ontology, with panprotopsychism positing the

existence of the extra element of protophenomenal properties. If Chalmers cannot construct such an argument for panprotopsychism, and we've seen that zombie thought experiments don't do it for him, then there is simply no reason to accept it over a purely cognitive alternative.

I am not trying to argue that panprotopsychism is necessarily wrong – as physics continues to develop and new properties of matter are discovered, I think that it is entirely conceivable that we could uncover some non-cognitive property of matter that lines up with the proto-phenomenal properties that Chalmers talks about. In the absence of any kind of scientific proof, however, I believe that the most plausible default position is that of the cognitivist. The onus is on the anti-cognitivist to explain why we ought to think that consciousness is *more* than just cognition; that it is somehow distinct from the rest of the operations of the brain.

The Mind-Body and Mind-Mind Problems

I have so far only mentioned the mind-body problem, and at that only briefly and indirectly. Quite simply, the center of the *mind-body* problem is the problem of how brain structure relates to cognitive function and consciousness (what Chalmers refers to as the easy problem of consciousness). This is a problem that is shared by cognitivists and anti-cognitivists alike – cognitivists believe that a completed neuroscience or cognitive science will provide us with all the answers about the nature of this relationship, where anti-cognitivists believe that science will never be able to provide a full account of this relationship, and propose non-cognitive entities as a means of filling this gap.

The introduction of non-cognitive entities is problematic, though, because it brings to light the related *mind-mind* problem: the question of how the conscious part of the mind interacts with the rest of the mind (Jackendoff 1987). The anti-cognitivist proposes that there are non-cognitive elements to consciousness, and as such needs to provide an account of how non-cognitive mental states or properties could possibly have a causal impact on cognition. If the natural world is causally closed (as all the available evidence suggests), then non-cognitive entities should not be able to enter into causal relationships with the rest of cognition[3].

Anti-cognitivists inherit this problem from dualists, as they are both proposing causal relationships between entities of different kinds. Descartes had an easy way out of this issue since he could make an appeal to the "divine". Clearly god is not limited by the laws of physics, and since Descartes' "souls" are meant to be on a similar plane of existence to god, the causal closure of the physical world is an issue that can be explained away through an appeal to omnipotence. He proposed that souls relate to bodies through a particular part of the brain, the pineal gland, but he could just as easily have claimed that the connection happens through communication between the soul and the heart, or the big toe. If you're hypothesizing about the powers of an omnipotent non-physical being, you have a fair bit of flexibility in terms of how you get around conceptual issues.

Contemporary anti-cognitivists do not have the same kind of flexibility, since belief in god tends to be a minority position in academic philosophy. Instead, anti-cognitivists appeal to various flavours of epiphenomenalism, or claim that the non-cognitive is in

[3] I should note that, although the mind-mind problem is also an issue for cognitivists, it does not pose the same kind of metaphysical problem for them as it does for anti-cognitivists, because the cognitivist ontology does not include non-cognitive entities/substances/properties.

fact a property of the cognitive that we do not yet have a full understanding of - an example of the latter approach being Chalmers' panprotopsychism. The issues with epiphenomenalism are huge, and the issue with Chalmers' approach is that that the mind-body problem and the mind-mind problem cannot be resolved, because panprotopsychism would involve kinds of entities that are not countenanced by contemporary physics and that do not have a clear place in the laws of nature. If the physical world is cognitively closed, then it would not be open to any intervention from the protophenomenal, effectively making Chalmers' approach a non-starter.

Beyond this issue of interaction between entities of different kinds, however, there is also the issue of interaction amongst non-cognitive entities. Suppose we grant the anti-cognitivist that consciousness is, in some way, non-cognitive. Presumably, there would be some kind of interaction going on here between the non-cognitive entities that consciousness is made up of. Especially on a view such as panprotopsychism, where consciousness comes from interaction among the non-cognitive properties of cognitive substances, there is probably some kind of causal relationship happening between "things" at the non-cognitive level. How does the anti-cognitivist suppose that this actually happens? Jaegwon Kim has given a similar critique of traditional Cartesian Dualism and the concept of soul-soul causation:

> The more we think about causation, the clearer becomes our realization, I think, that the possibility of causation between distinct objects depends on a shared spacelike coordinate system in which these objects are located, a scheme that individuates objects by their "locations" in the scheme. Are there such schemes other than the scheme of physical space? I don't believe we know of any. This alone

> makes trouble for serious substance dualisms and dualist conceptions of personhood… (Kim 2001, p.43)

Kim argues that not only are the possible mechanisms behind non-physical to physical causation unknown, but even trying to imagine what causation between non-physical entities could look like is a difficult task. Kim suggests that, although it may be possible for non-physical entities to be assigned some sort of non-physical spatio-temporal coordinate system, we have no way of knowing how such a coordinate system might work, or even how to go about thinking about what one might look like (p.37). How would various "parts" of the conscious mind communicate, or have any kind of causal relationship? As Kim, I think rightly, points out, causation seems to necessarily entail physicality due to the need for spatiotemporality in relations between things. This renders consciousness understood as a non-cognitive entity unsupportable, since it is both unclear how causation might happen at the non-cognitive, non-spatiotemporal level, and therefore how consciousness could have any kind of role in guiding action, since it has no potential for causation. Chalmers' brand of anti-cognitivism faces exactly the same difficulties as traditional substance dualism does. If there's no place for the non-cognitive in the laws of nature, and if the material world is cognitively closed under the laws of nature, then there is no possibility of causal interaction between the panprotopsychic and everything else.

The panprotopsychist could reply that claiming that a non-cognitive entity is made up of "parts" is to wrongly apply our traditional physical understanding of the universe to something wholly different. Perhaps the non-cognitive is some sort of simple force, not composed of parts and not similar or analogous to the cognitive in any way.

But the issue with such an approach is that it would be ad hoc, and perhaps even unfalsifiable – the anti-cognitivist in theory can always find a way out of a logical corner with this kind of move. Since the non-cognitive is completely unknown to us, it is of course possible to hypothesize the existence of some entity that conveniently avoids all of the critiques of the cognitivist. However, since we have no reason in the first place to believe in the non-cognitive (as per Occam's Razor), and have seen no evidence whatsoever to support any conclusions about the nature of proposed non-cognitive substances, we are once again left with no reason to accept such claims. The ability to conceive of some substance that neatly explains the phenomenon that we are examining does not give us license to accept that such a substance exists.

1.3 - Conclusion

In this chapter I have not tried to give a particular cognitivist explanation of consciousness that we ought to make use of. The details of unravelling the inner workings of the mind is an empirical matter - a job perhaps better suited to neuroscientists and cognitive scientists than philosophers. All that I have tried to do with this chapter is establish the ontological perspective from which I believe we ought to approach consciousness research, and the one that I will assume throughout the rest of this thesis. This perspective opens the door to a wide variety of different pictures of consciousness (higher-order thought, first order representation, etc.), and does not impose arbitrary limitations on what kinds of creatures we take to have conscious experience, such as the necessity for consciousness of self does.

As I noted earlier in this chapter, it is not impossible that there exist non-cognitive properties of consciousness, however in the absence of any compelling argument, I see little reason to accept such a view. With this issue in mind, I do not find Chalmers' critique of cognitivism as "incoherent" (relative to anti-cognitivism) to hold much water. Indeed, I hope to have demonstrated that anti-cognitivism is, in all of its suggested forms, a decidedly confused and incoherent view in its own right. As Occam's Razor suggests, the burden of proof lies on the anti-cognitivist to explain why we ought to accept non-cognitive substances into our otherwise pleasingly simple ontology. Thus far, I see no compelling arguments in defense of such a position.

Chapter 2: Between Two Horns

So far I have argued that consciousness, whatever it is, should be treated as a purely cognitive phenomenon. We have no evidence in favor of taking a non-cognitive approach (other than a supposed lack of evidence in favor of a cognitive approach), and Occam's Razor provides a compelling reason to resist adding non-cognitive substances into our ontology. Still, this tells us very little about what consciousness actually is - asserting that a leg is part of a human body tells you roughly where legs might be found, but it does not tell you much about how legs work, or what exactly legs are constituted of. In this chapter, I will begin to explore the question of what characterizes consciousness beyond its ontological foundation.

2.1 - What "Consciousness" Refers To

Something that almost all theories of consciousness have in common, regardless of the details of each one, is that the difference between a conscious thing and a non-conscious thing is the presence of experience. One of the reasons that Nagel's approach has left such an indelible mark on the literature is because it is able to convey, in an intuitive and evocative way, what it means to be a conscious thing. When asked to explain the difference between the way that I experience the world and the way that a rock experiences the world, the most fundamental difference is that it is *like* something to be me, and it is not *like* something to be a rock. It is not, for instance, that the rock doesn't have language, or sense organs, or a conception of time, since all of these characteristics could in

principle also be realized by a non-conscious system. What could not be realized by a non-conscious system, though, is experience.

Furthermore, since a system that is experiencing anything is by definition conscious, and a system that is conscious is by definition experiencing something (the flip side of this being that without experience, there can be no consciousness, and without consciousness, there can be no experience), I take experience to be both necessary and sufficient for consciousness (though I acknowledge that this may be to some extent stipulative). As such, when I refer to a conscious system, I am referring minimally to a system that has experience.

Another way to frame this idea is in terms of Gottlob Frege's "sense/reference" distinction. Frege asserted that there are two kinds of information conveyed within a noun, the sense and the reference (or *bedeutung*):

> A proper name (word, sign, combination of signs, expression) expresses its sense, stands for [bedeutet] or designates [bezeuhnet] its Bedeutung. By employing a sign we express its sense and designate its Bedeutung. (Frege 1892, 32)

Frege used the example of the planet Venus to illustrate his point. In antiquity, Venus was known as "Venus", "The Morning Star", and "The Evening Star". At the time, it was not known that each of these terms referred to the same celestial body – it was assumed that the star that was visible in the morning (Venus) was not the same star that was visible in the evening (also Venus). In this case the reference for all of these proper names is the same, because they all refer to the same celestial body. However, the sense is different because the sense concerns the mode of presentation in which information is

conveyed. As a result, if you are unaware that the Morning Star and the Evening Star both refer to Venus, then you would not understand the names as being interchangeable. Applying this line of thought to the concepts of consciousness, experience, and phenomenology, I might suggest that what we have here is difference in sense, but not a difference in reference. I take these terms to refer to the same phenomenon, just in different modes of presentation.

One of the obvious difficulties with what some might call *definitions* of consciousness (or experience, or phenomenology), is that it is very difficult to provide such a definition without being uninformative and circular. A common critique raised against Nagel is that his "what it is like" language does not seem to tell us very much about what consciousness is. It is evocative, to be sure, but it doesn't make a substantial contribution to our understanding of consciousness. If pressed to explain what it means for there to be "something it is like to be *x*", at some point Nagel would likely have to make reference to a term like awareness, wakefulness, experience, or some similar concept. Unfortunately, defining what any of these concepts mean will eventually require an appeal to the concept of consciousness itself. This is, of course, a sub-optimal way to define a concept. If you don't know what consciousness is in the first place, then a circular definition that loops back to consciousness at some point will not be useful to you. The most that this method of jumping from a concept to a near-synonym can contribute is that it might help us to narrow in on the general idea behind the concept that we're trying to articulate and generate examples of consciousness that can get general inter-subjective agreement. Agreeing on what constitutes an example of consciousness is certainly helpful, because the more examples we can agree on, the more accurate our evaluation may be of the common

characteristics shared between them. However, it doesn't get us much closer to a proper definition of consciousness.

In an effort to avoid this issue of circularity, some theorists have chosen to go in the opposite direction from those who concern themselves with defining the elusive and slippery concept of consciousness. These theorists instead focus on giving an explanation of what it means to be conscious through an appeal to the cognitive functions that are likely to underlie it. This approach allows for an explanation of consciousness of the sort "consciousness (whatever it is) occurs when *x* and *y* cognitive function occur". Rather than focusing on the experiential nature of consciousness, these theorists approach consciousness through an appeal to particular cognitive states, allowing for a non-circular, in some cases reductive, explanation of consciousness. Bernard Baars' Global Workspace Theory (GWT) is one example of such an approach. Baars is primarily concerned with the function of consciousness within the context of cognition in general, and his theory posits that consciousness, or as he puts it the "Conscious Awareness System", acts as a kind of mediator between different parts of the brain. Baars writes,

> [GWT] suggests a fleeting memory capacity that enables access between brain functions that are otherwise separate. This makes sense in a brain that is viewed as a massive parallel set of specialized processors. In such a system, coordination and control may take place by way of a central information exchange, allowing some processors — such as sensory systems in the brain — to distribute information to the system as a whole. (Baars 2005)

On this picture, consciousness is presented as facilitating the global broadcast of certain pieces of information that are made available to different cognitive systems. Baars illustrates GWT with a theater analogy – consciousness is like a spotlight shining on a stage, where everything within the narrow scope of the spotlight is conscious, and everything left un-illuminated is unconscious. Thinking of consciousness in this way, as a director of attention, provides a theory that can be scientifically evaluated. For instance, Baars cites a study by Morris Moscovitch that tracks the way that visually gathered conscious information is distributed and made available to unconscious parts of the brain, as providing empirical support for his hypothesis (Baars 2005).

Baars' explanation of the cognitive function of consciousness is not circular, but it could be argued that it *changes the subject* from consciousness to perhaps only loosely related cognitive functions. If such an argument is correct, then what Baars and others like him are really doing when they try to discuss consciousness is a kind of sleight of hand – they move from consciousness to cognition without providing a plausible answer as to how and why cognitive functions give rise to consciousness (Chalmers' hard problem).

Nagel's and Baars' very different approaches highlight the two most common ways that articulations of consciousness, or experience, can go wrong. I understand their two very different approaches as representative of the two horns of a dilemma. When we try to *define* consciousness, we are limited to doing so through appeals to near synonyms, and as a result we run into the horn of *circularity*. When we try to give an explanation of *what it means* to be conscious through appeals to cognitive functions, and do so without providing a plausible answer to the hard problem, we run into the horn of *subject changing*. The issue here is that there are elements of both horns that must be included in an accurate articulation of what consciousness is (unless you are an anti-cognitivist and believe

that you already refuted that idea). We need to address phenomenology, but we need to do so in a way that explicable in cognitive terms, or else risk venturing into the murky waters of anti-cognitivism. In the following sections, I will provide more detailed accounts of the problems of these horns, and explore whether it is possible to retain the important elements of both horns and shed the problematic ones.

Horn 1 – Circularity

The primary way that *definitions* of consciousness go wrong is by lapsing into circularity. If one tries to define consciousness by reference to similar terms such as wakefulness, awareness, attention, point of view, what-is-is-like-ness, and so on, one will eventually have to complete their circular explanation by making reference to consciousness. This is a problem that is not unique to consciousness, as we also run into circularity when trying to define other, similarly fundamental, concepts. Consider for instance the concept of "matter". Suppose you are tasked with defining the term "matter" in a way that would make sense to a being that has no concept of what matter is. What you will quickly realize is that every attempt you make to define matter falls into the same kind of circularity. The conversation is likely to go something like this:

A: Matter is the physical substance from which everything in the universe is made.

Q: Okay, but what is a physical substance?

A: Well, it's material.

Q: And material is…?

This could go on forever, with both parties growing increasingly frustrated with the incompetence of the other. Matter is one of these interesting concepts that seems to be indefinable without resorting to some kind of circularity. We can give examples of matter – an apple, a space rock, a molecule – but reaching an agreement on what counts as an example of some concept can only get us so far in our attempts to understand it. We could also try to explain it by becoming increasingly specific in our explanations – matter is made of molecules, which are made of atoms, which are made of protons, which are made of quarks, which are made of (…), but once again all that we're doing is giving examples of what matter is. Eventually we will reach a point where we can go no smaller, and at that stage we will be forced to say that whatever microscopic thing we're talking about is simply "matter". In instances where we are trying to provide definitions of fundamental concepts such as matter or consciousness, we can expand the explanatory circle, but it is perhaps impossible to get away from the circle entirely.

A recent attempt at defining consciousness that I take to be unsuccessful at avoiding this horn is Michael Tye's "Simple View". Tye articulates the Simple View in the following way:

> [A] creature is conscious at time *t* if and only if it is undergoing one or more experiences at *t*. Further and relatedly, a mental state is conscious if and only if it is an experience. (Tye 2017, p. 12)

Tye's view is clearly influenced by Nagel's approach, and I think that his use of the phrase "undergoing one or more experiences" can be understood as a modified way of phrasing Nagel's "something it is like" explanation of consciousness. Saying that a

creature has an experience is just another way of saying that there is something it is like to be that creature – it doesn't tell us anything substantive about what exactly experience is, or whether experience and consciousness are the same, or just related concepts. Tye's definition is also uninformative – he explains that a creature is conscious if it is undergoing an experience but doesn't make explicit what he thinks experience actually is.

Tye's approach falls victim to the first horn of the dilemma in that it is unabashedly circular. Defining consciousness simply as experience is useful in the sense that it can help to make clear what it is that differentiates consciousness from a lack of consciousness, but with minimal probing it quickly degenerates into an endless back and forth cycle of defining consciousness through appeal to experience, and vice versa. Claiming that "a mental state is conscious iff it is an experience" may be true, but on Tye's account consciousness and experience are taken to be equivalent, and since he provides no extra insight as to what experience is, such a claim doesn't tell us very much. This makes Tye's account about as informative as claiming that "a mental state is conscious iff it is conscious".

Despite the issues with Tye's approach and others like it, in order to articulate what consciousness is, we clearly need to provide an account of its phenomenal nature. As a result, any approach to articulating consciousness is going to include at least *some* circularity. Where this tends to go wrong is when the phenomenal is the beginning and end of the discussion – when no cognitive explanations are provided as a means of getting at what constitutes experience. Cognitive explanations allow for an easy exit from this circularity because they articulate consciousness through an explanation of something that is not itself conscious (cognition). In response to the circularity of theories like Tye's Simple View, some theorists have chosen to put defining experience on the backburner by

focusing instead on explaining what kinds of cognitive functions are likely to underlie it. This approach is not without its own drawbacks, though, as it opens one up to the problems associated with the second horn of the dilemma.

Horn 2 – Changing the Subject

The primary way that *explanations of the function* of consciousness tend to fail is by trying to explain consciousness through an appeal to a particular cognitive function or group of functions that could be non-conscious, without providing a plausible bridge over the explanatory gap (the 'gap' between the physical and the phenomenal). Explanations of this sort typically have their roots in the findings of cognitive science, and so have a much stronger empirical backing than those that come from the armchair analysis of folk psychological concepts that are so common amongst approaches that take phenomenology seriously. A general critique raised against these sorts of explanations is that they "change the subject" – they aren't actually talking about consciousness, but instead some cognitive function that, despite perhaps playing a role in consciousness, is not sufficient for consciousness. Cognitive states are undoubtedly what states of consciousness consist in, and as such any approach to the study of consciousness ought to be in agreement with cognitive science, but by focusing only on cognition without showing how cognition provides an account of experience, accounts of this sort fall short of providing a plausible explanation of the mechanisms behind consciousness. On the subject of how approaches to theories of mind have evolved over centuries of study, Tononi writes,

> Materialism, or its modern offspring, physicalism, has profited immensely from Galileo's pragmatic stance of removing subjectivity (mind) from nature in order to describe and understand it objectively. But it has done so at the cost of failing to deal with the central aspect of reality – experience itself. (Tononi & Koch 2014)

Although I find Baars' hypothesis about the role that consciousness plays in cognition to be quite plausible, it is not difficult to see how his theory falls victim to this horn. Once Baars moves from discussing experience to discussing the ways that different cognitive functions facilitate the communication of information, he has turned to something that need not be conscious, without telling a plausible story about how these mechanisms actually give rise to experience. It seems plausible that a cognitive system could have all of the things that Baars connects to his "conscious awareness system" – such as centralized information processing and global distribution of information – without actually having experience. The root of the issue here is a kind of interface problem – explaining how experience can arise out of processes that individually have no phenomenal quality is notoriously difficult, as shown by Chalmers' investigation of the hard problem and the huge amount of ink spilled over the explanatory gap between cognition and consciousness. Put another way, it is very difficult to isolate particular cognitive functions that *necessarily* give rise to, or are accompanied by, experience. The result of this issue is that appeals to cognition often fail to provide an adequate explanation of what consciousness is.

Interestingly, despite changing the subject, Baars' approach also seems to run into the circularity issue present in most definitions, since it is unclear on his theory how to carve up the distinction between consciousness, attention, and awareness. When pressed to explain consciousness as directly as possible, Baars writes:

We can say that mental processes *are conscious* if they

(a) Are claimed by people to be conscious; and

(b) Can be reported and acted upon,

(c) With verifiable accuracy,

(d) Under optimal reporting conditions…

(Baars & McGovern 2007)

Obviously, this is not really a definition of consciousness *per se*, but more a way of classifying some mental states as being conscious or unconscious (for an argument against the value of reportability, see Chapter 3). Baars compares consciousness to other scientific concepts such as "heat", in that "formal definitions…tend to come quite late, often centuries after adequate operational definitions are developed" (Baars & McGovern 2007). Perhaps this is a fair comparison, but I think that there are important differences between the two concepts. Heat has the advantage of being able to be empirically detected, observed, measured, and its effects seen. What makes consciousness uniquely difficult to study is that it seems to lack all of those qualities, and as a result, we are unlikely to make the same kind of progress as we do with other scientific concepts. When undertaking the study of heat, one is unlikely to inadvertently start studying trees, or rocks. When undertaking the study of consciousness, however, it is very easy to slip into the study of potentially irrelevant cognitive functions. To use Baars himself as an example, it seems entirely possible to have (a) claims of consciousness, (b) reports and actions reflective of those claims, (c) verifiable reports, and (d) optimal reporting conditions, *in the absence* of consciousness, and vice versa because you can have consciousness without any of those conditions being met.

One need only consult the voice assistant in their smart phone to see why this is the case. It is therefore of particular importance in the case of consciousness that we work towards making our conceptual toolbox as neat and orderly as possible, or otherwise risk studying only loosely related phenomena.

What Can Be Retained?

Although the horns of the dilemma pose serious issues for articulating what it means to be conscious, we cannot simply avoid the horns by doing away with them altogether. We need to give an account of the phenomenal nature of consciousness, and this puts us at risk of circularity, but we also need our account to be cognitive, which puts us at risk of subject changing. Furthermore, and importantly, we need to provide a plausible link between cognition and experience – one that crosses the explanatory gap between the two.

One way to approach this problem is by identifying the kinds of cognitive functions that are necessary for experience. In most cases that I have seen, theorists who attempt to provide a cognitive account of consciousness make the mistake of proposing cognitive functions that, despite being tangled up with consciousness, are not strictly speaking *necessary* for it. For instance, it seems plausible that consciousness typically entails some form of memory, as memory is a necessary part of belief formation, verbal reporting, and comprehension of one's environment – features that, depending on which particular account of consciousness you are operating with, may seem more or less important. However, it certainly seems possible for there to be experience in the absence of memory, because all that a mental state requires to be an experience is that it be like something to

have that mental state. The same can be said for mental states that, to use Baars' language, are "claimed by people to be conscious" – if this is the bar that we set for a determination that some mental state is conscious, then we are likely to let even unsophisticated forms of AI into the class of conscious beings.

If we take experience as necessary and sufficient for consciousness, then we can focus on trying to articulate what it means to have an experience and in doing so also provide an account of the phenomenal nature of consciousness. This would allow us to avoid circularity, because we can say that what makes some system conscious is the ability to have experience, but also show that experience is constituted by *x cognitive function.* This appeal to experience allows us to take advantage of the useful kind of example-generating language that (albeit circular) definitions like those of Nagel and Tye elicit, and the appeal to cognition allows for a means of escaping the circularity. Of course, this still leaves the link between cognition and consciousness to be forged, so the kind of cognitive function that will need to be proposed must be one that directly and necessarily gives rise to experience if we are to avoid changing the subject. I believe that one promising way of achieving this is through an appeal to *mental representation.*

2.2 - The Minimal Claim

One way to avoid the problems associated with these horns is to start from a very general idea – that consciousness has a *representational base.* This I will refer to as the Minimal Claim (MC). Before discussing the link between representation and experience, I will first lay out what I mean when I use the term "representation".

What is Mental Representation?

By "representational", I am referring to the phenomenon of mental representation; internal mental states that are generated to represent internal and external states. When I stare out my window, my visual system transmits information about my environment (the orientation of the objects in my field of view, their colour, the amount of light that is bouncing off of them, their texture, etc.) to my brain, which then uses this information to create a mental representation of my environment. Mental representations need not be generated only by sense perceptions though, as the information that they are generated from can have an internal source – visual and auditory hallucinations are extreme examples of such a phenomenon.

Mental representations are commonly characterized as having two primary features. The first of these features is *truth evaluability* – representations are mental states that are truth evaluable, in that they either accurately represent some feature of one's environment or internal states, or they do not. The reason that we can identify when someone is hallucinating, for instance, is because we can determine that the content of their mental representations does not correspond with real events in their environment. The second of these features is that mental representations can act as a *surrogate* for actual perception – the reason that I am able to think about my cat without actually perceiving my cat is because I can conjure up a mental representation to serve as a surrogate for perceiving him. This surrogate function underlies a great many cognitive capacities, such as memory and pre-planning of action, as it allows one to engage with past and potential perceptions.

Representation and Experience

I take 'consciousness' and 'experience' as referring to the same kind of thing, however, neither is interchangeable with representation. By proposing that consciousness has a representational base, I am proposing a part/whole relationship, with mental representation constituting a necessary condition for consciousness. I take this claim to be plausible for two reasons. On one hand, to make the claim that consciousness is equivalent to representation would require the dubious and widely-disagreed-with view that all representations are conscious. It is clear that there is a subset of representations that are unconscious, and so, such an equivalency would not be correct. On the other hand, though, it is clear that there is also a particular kind of mental representation that underlies consciousness, in that conscious content is representational content.[4]

I take the MC to be a relatively uncontroversial claim, as representation is perhaps one of the only common threads across cognitive accounts of consciousness. If we take representation as being a necessary condition for consciousness, then this leads to the question of what it is exactly that constitutes a *sufficient* condition for consciousness. If there can be both unconscious and conscious mental representations, then there must be some difference between the two – some process that conscious representations have undergone that unconscious ones have not. Various forms of representationalism have involved attempts at explaining this process, with perhaps the two most influential

[4] I do acknowledge that there are some theorists who deny the existence of representation, though I take them to occupy a minority position and therefore will not address them in much detail here. I will note though that typically theorists of this sort propose something in the place of representation that nonetheless serves the same function, and it could be argued that in so doing they are simply reintroducing representation under a different name.

approaches being Higher Order Thought (HOT) theories and First-Order Representationalism (FOR). HOT theorists posit that a mental representation is made conscious when it is the target of a higher-order thought. On this picture, the feeling of pain is made conscious only when a corresponding higher order thought is directed at the pain. Proponents of FOR argue that phenomenally conscious states need not be the target of a higher-order thought, and that it is enough by proponents of FOR for a state to be accessible to "lower" levels of cognition for it to be phenomenally conscious (though the details of how this happens tends to vary greatly between proponents of FOR). One of the more influential accounts of FOR is Michael Tye's PANIC theory (a state is conscious when it has Poised, Abstract, Non-Conceptual, Intentional Content) (Tye 1995).

These theories are not without their problems, though. FOR approaches commonly struggle to account for what consciousness, and consciousness of self, is like. Tye's PANIC theory tells us what kinds of characteristics a conscious mental representation has, but not much about the mechanisms that facilitate their having these characteristics, or how/why these characteristics *result* in experience instead of a lack of experience. In principle it seems possible for a state to have all of the aspects of a PANIC state, and not be conscious. As a result, Tye's account of FOR (and others like it) struggle to provide a bridge over the explanatory gap, and are therefore also guilty of subject changing.

HOT theory also has serious shortfalls, one of which being what is sometimes referred to as the "empty HOT" objection (Block 2011, P.424). It seems that HOT runs into problems when it is applied to highly abstract thought that lacks first-order representation – for instance it is not clear that how a higher order thought might be involved in *imagining* oneself doing something. If there is no lower-order representation (such as in the case of imagination, or a feeling of contentedness directed at nothing in particular), then

there is nothing to direct a HOT at, and if there is nothing to direct a HOT at, then there should be no consciousness, and yet there is. The issue here is that HOT theory lays out necessary and sufficient conditions for a state being conscious, and yet we can find examples of conscious states that do not meet these necessary and sufficient conditions. Block has referred to this as the "fatal objection" to HOT theory (P.420). A related issue is that we seem to be conscious of things that we are not having any sort of higher-order thought *about* – I can simply be aware of the can of sparkling water on my desk without having a higher-order thought about it. Although the empty HOT objection may provide reason to doubt HOT theory, there have been some strong responses to it that I will not go into here (such as from Rosenthal and Mylopoulos). Regardless, how plausible one might find the MC ought to have little to do with the mechanics of how representations are made conscious, and as such, I will not take a stand on the issue here.

A further issue, and one of particular relevance to my project, is the issue of how HOT theory applies to cognitive systems that do not have the capacity to have HOTs. Infants, non-human animals, and neurologically atypical adults may all be candidates for such a deficiency, but in some cases withholding an ascription of consciousness from these organisms would not be consistent with what we know about them. On this issue, Clark writes,

> The immediate worry about the higher order approach is that it seems to tie phenomenal consciousness to the presence of other advanced meta-cognitive capacities. Thinking about your thoughts is, on the face of it at least, something that most animals and young infants are probably unable to do. (Clark 2001)

Clark goes on to assert that proponents of HOT theory need to either accept the claim that many organisms including all non-human animals are not phenomenally conscious, or weaken the definition of a higher order thought to the point that it could plausibly occur in organisms other than humans. An argument of the latter kind would probably involve a portrayal of HOTs as being not necessarily conceptual, or in Clark's words, *rationalistic,* though it is unclear whether such an approach could retain the proposed role that HOT's are intended to serve. Currently there is little agreement amongst proponents of HOT theory as to which compromise to make.

One approach to sufficiency that I think is particularly promising, in that it skirts the problems associated with HOTs and FORs, involves the notion of self-presenting representations (SPR). Proponents of SPR suggest that it is not the presence of a secondary, higher order thought that makes a representation a conscious one, but instead a feature of the representation itself. SPR is a well-established theory, and there are various accounts of how self-presentation occurs that I will not explore (Kriegel, Wider, Williford, Van Gulick). One approach that I do think is worth mentioning though is that of Brook (2006), who proposes that conscious representations may be those that include three different kinds of content:

1. The content of the representation (the green-ness of the leaves)
2. Information about the representation (clarity of one's vision)
3. Awareness of the subject of the representation (the self as experiencer of the visual perception)

Contrarily, a representation could fail to be a conscious one in virtue of lacking one or more of these kinds of content. For instance, simple phenomenal consciousness could not lack (1), because for there to be something it is like to have an experience, the experience needs to have content. Consciousness of self could not lack (3), because for there to be consciousness of self, there of course needs to be awareness of the self. Further to this idea, it may be possible for a system to have only the capacity for representations that lack all three kinds of content, making such a system non-conscious. Brook makes the important point that consciousness of the self as experiencer does not necessarily entail *knowledge* of self, at least not the conceptual, abstract kind of knowledge that we would normally associate with such a term (see Shoemaker 1968 "self-reference without identification"). Additionally, one's agreement with the existence of the second kind of content listed above will be largely dependent on how one views the issue of transparency, because if representations are in fact transparent, then being aware of the representation itself is impossible. As I have discussed so far, I do not think this is an issue for SPR, because we have good reason to believe that representations are not transparent. I take the notion of SPR as being the most plausible approach, as it seems to not have the major issues that are associated with the alternatives of FOR and HOT.

Whichever account of the mechanisms of conscious representation that we choose to subscribe to, though, one thing that is clear is that since representation is universally necessary for consciousness, it will be part of any sufficient condition. The MC establishes the necessary cognitive foundation for consciousness, upon which further inquiries can be made regarding which other cognitive functions or properties are sufficient for consciousness (it seems to me that SPR would fill this role, but any account of the relationship

between representation and consciousness is a candidate for having identified what the sufficient conditions are).

The Minimal Claim and the Horns

The MC is not a circular claim, as it does not try to draw an equivalency between consciousness and representation, but what about the second horn – subject changing? I anticipate two main criticisms against the MC – one could argue either that it does not escape the charge of subject changing because there could be representational states that are not conscious, or that it successfully escapes the charge of subject changing but at the cost of a picture that makes representation so unlike the rest of cognition that it just repeats the mystery without reducing it. In the latter case, it could be argued that representation is being used as a kind of magical do-it-all concept that, despite providing answers to our questions, becomes so vague that it does not carry the argument much further.

I will address the latter criticism first. The MC does not require that representation is an immaterial, proto-phenomenal, or similarly mysterious non-cognitive property or process. Representation is a cognitive function that is similar to others, like memory formation and recollection, visual perception, belief formation, and so on. In other words, there is nothing spooky about mental representation, and as a result the MC is entirely compatible with representation being cognitive. Indeed, representation is a widely recognized cognitive capacity – one can find reference to it in the fields of psychology, neuroscience, cognitive science, and philosophy, so the notion that it is in some way deeply mysterious would not be justified.

Many cognitive accounts of consciousness get around circularity by proposing a cognitive capacity that is actually so far from consciousness that they begin to discuss something that appears to be completely different. I think that the MC is unlike such accounts because it establishes a necessary condition for consciousness, and not anything else, and provides a foundation upon which accounts of sufficient conditions for consciousness can be built (through appeals to FOR, HOT, SPR, etc.). This gives it an advantage over other approaches that propose explanations of consciousness that include cognitive functions that could happen in the absence of experience, or that experience could happen in the absence of. In this way, the MC is far closer to giving an account of phenomenology than competing cognitive accounts and is able to evade charges of subject changing.

2.3 - Conclusion

Representation is foundational to almost all contemporary theories of consciousness, anti-cognitive or cognitive. As a result, the Minimal Claim should find wide agreement amongst philosophers of mind and cognitive scientists alike. In addition to being relatively uncontroversial, the MC is uniquely well formulated to retain what is valuable in both horns of the dilemma and does not suffer from the pitfalls that so many other theories of consciousness do. It is able to keep the focus on phenomenology and take into account the findings of cognitive science, without falling victim to circularity or subject changing. Furthermore, the MC does not propose any feature of consciousness that is not already part of cognitive science, protecting the MC from charges of introducing mysterious entities that no science could ever deal with (such as in the case of the mysterianism of

Chalmers and Block) and improving its chances of being consistent with future scientific findings.

Finally, I believe that the MC can provide a strong foundation from which we can take on long-standing problems in philosophy of mind, particularly the question of how we might go about determining whether beings other than ourselves, human and/or non-human, are conscious. If we take a particular kind of representation as being necessary and sufficient for consciousness, then the identification of such a representation in another being would give us conclusive reason to believe that said being is conscious. In the following chapter, I will explore the relationship between mental representation and behavior, and discuss to what extent behavior can provide sufficient conditions for establishing whether some system is conscious.

Chapter 3: Representation and Behavioral Flexibility

If you ask most pet owners about whether they think their pet is conscious, I expect that the typical response would be a confident "yes" – it seems intuitively obvious that our animals are conscious creatures. When my cat meows at me at around 8:00 in the morning (his breakfast time), I am naturally inclined to ascribe to him certain mental states. I assume, for instance, that he is feeling hungry, that he intended to get my attention by making noise, that being hungry is an uncomfortable experience for him and that he would like it to end, that he believes in one way or another that I am capable of providing food for him, and so on. In fact, it would be quite odd for me to think that my cat isn't feeling hungry, but rather that my cat's stomach has emptied and this has resulted in a signal being sent to the brain, which triggered a behavior intended to contribute to the filling back up of the stomach, all entirely unconsciously.

This kind of explanation would be likely to raise the eyebrows of people who are unfamiliar with the debates surrounding consciousness and unreportability, because it does not capture what our intuitions tell us is happening. For those who are familiar with such debates, such a description might be quite pleasing, in that it avoids ascribing experience to a creature that could lack it altogether. While I think that most consciousness researchers share the intuition that at least some animals are conscious, it is difficult to justify our intuitions, and these intuitions are at times challenged by our reliance on verbal reporting.

In the following chapter, I will attempt to provide some justification for our intuitions through an appeal to mental representation and behavioral flexibility, and will demonstrate through a case study on tool use how an approach that factors in both of

these elements can be practically applied in an evaluation of externally observable behavior. I should make clear from the very beginning that I do not intend to present my argument as being definitive – it is conceivable that even the most suggestive behaviors could happen in the absence of conscious experience. Rather, I hope to demonstrate that mental representation and behavioral flexibility provide some promising footholds that we can use in determining whether some animal is at least very likely to be conscious.

3.1 - What Are Indicators?

Indicators vs. Criteria for Consciousness

There are two possible approaches to interpreting behavior and cognitive features and what they can tell us about consciousness. The first approach, which I will refer to as the criterion approach (CA), is to consider some feature a *criterion* for consciousness, meaning that without this feature, a creature cannot possibly be conscious, and with this or some other feature, a creature cannot *fail* to be conscious. This model has some intuitive appeal, because it makes ascriptions of consciousness very straight forward – all we need to do is match the features of some system to a checklist of necessary and sufficient features for consciousness, and if all the boxes are ticked on our checklist, we are dealing with a conscious creature. A notable proponent of the criterial approach is P. F. Strawson, who argues that there are observable behaviors that constitute adequate criteria for an ascription of "P-predicates" – a class of predicates that include mental states (Strawson 1959, p.105).

At this point in time, I do not believe that the CA is of much practical use. Beyond what I have argued, that consciousness has a representational base, there is little more that can be said with confidence about consciousness. To take a criteria approach would be to adopt an unobtainably precise method of determining the presence of consciousness, given what we actually know about the topic. There is such a great degree of variation in behavior in the world that, even if we could find behavioral criteria that applied to all behavior of a given kind of conscious system (humans, for instance), it is unlikely that such criteria could be neatly applied to other kinds of conscious systems (non-human animals).

The indicator approach (IA) is, as the name suggests, less stringent. Instead of looking for hard requirements for an ascription of consciousness, the indicator approach involves looking for behavioral features that *indicate* the likely presence of consciousness. Indicators have the advantage of adding significant flexibility and inclusivity (within the limits of what we can actually do) to our ascriptions of consciousness – rather than saying for example "*x* system must have the capacity to report on mental states to be considered conscious", on the indicator approach we can say "if *x* system is capable of reporting on its mental states, the likelihood of *x* system being conscious increases". The IA is of particular use when applied to cognitive systems that are different from the average healthy adult human, because it can quite naturally accommodate multiple realizability. It can be just as easily applied to an elephant, octopus, or a human in a persistent vegetative state, as to me. With indicators of consciousness, the more indicators we can identify, the higher the probability of some system being conscious, and the fewer indicators we can identify, the lower the probability of consciousness.

Given what little we know about consciousness, I take the IA to be a far more plausible approach than that of the CA. At this point in time we simply do not know enough about the behavioral features of consciousness to make any serious claims about what features are sufficient for consciousness. Furthermore, due to the seemingly limitless number of ways that conscious creatures can behave, it is not clear to me that we will ever be able to justify such a claim. As a result of this lack of certainty about the nature of consciousness, I take the IA to be the only reasonable option that we have, and it will be the approach that I take going forward.

Marian Dawkins articulates a related and important issue that she refers to as the "paradox of animal consciousness", where in order for the study of animal consciousness to provide conclusive answers, it must be approached as a science, but experience is one thing that it seems impossible, or at least very difficult, to study as a science. (Dawkins 2014, P.26). Despite this paradox, however, we do not hesitate to ascribe conscious mental states to animals. Dawkins suggests that addressing this paradox may involve simply accepting the inherent tension and uncertainty, and embracing the "next best thing" that is the probabilistic conclusion that results from a wholistic evaluation of indicators of consciousness. The best that any approach to animal consciousness can offer us is what can only be probabilistic conclusions, and with this in mind we ought to focus our energy on doing what we can with our probabilistic approaches, rather than dismissing them as wholly uninformative.

One of the criticisms of probabilistic approaches is that they appear to render knowledge of other minds impossible. If we never truly know whether our probabilistic assessments are accurate or not, then we can never know with certainty whether others are actually conscious. While I would agree that this is a potentially uncomfortable

conclusion of my argument for an indicator approach, I also think that it is correct. We have intuitions about whether others are conscious, and we can justify these intuitions through an appeal to probabilistic indicators, but there is no way to gain certainty about someone's experience (except maybe by *being* that person, if awareness of our own states counts as knowledge). The same is true for many other questions in epistemology – it could be argued that I can never know with certainty whether I am a brain in a vat, for instance. Despite this problem, most of us take well-justified if probabilistic judgments about the nature of knowledge as being *as good as* certainty in cases where certainty is unattainable. If knowledge of other minds is one such case, as I argue that it is, then the criticism that the indicator approach would make knowledge of other minds impossible would be mistaken.

Indicators vs. Constituents of Consciousness

A further distinction that should be made at this point is between indicators and constituents. Indicators of consciousness are, as I have described them, behavioral qualities that indicate a probability of consciousness. Constituents of consciousness are quite different; in that they would actually be what consciousness is constituted of. A claim about constituents of consciousness would take the form of "consciousness is made up of / constituted by *x, y, z,* processes / structures". One can identify as many behavioral indicators of consciousness in an organism as one likes, and they will still only amount to inconclusive support for an ascription of consciousness. However, if we were able to understand

what the constituents of consciousness are, and identify all of them in an organism, we would be able to justify a conclusion that the organism is conscious.

A useful example of this idea can be found in the world of medicine, because the way that we diagnose and understand disease is also guided by the indicator/constituent distinction. When a patient is presenting with symptoms such as a runny nose, sore throat, and malaise, a doctor is likely to diagnose a common cold. This is because the symptoms that the patient presents with are *indicators* of a particular virus. They are certainly not what the virus is, or what *constitutes* the virus, but we are nonetheless able to confidently arrive at a diagnosis. Another example of this is the way that we determine that the effects of anaesthesia have worn off after a surgery. What persuades us that a person has regained consciousness? It is that a rich repertoire of flexibly controlled behavior begins again. Patients begin to move, respond to stimuli, and engage in fine motor control, and this has been shown to reliably justify our conclusion that the person is no longer under the incapacitating effects of anaesthesia. In many cases, doctors are able to make accurate diagnoses based on the presence of indicators, rather than constituents, of conditions. Although symptoms on their own can and often are compatible with a great many ailments and provide little in the way of reliable information, the right combination of symptoms can often tell us quite a bit about what might be happening.

Acknowledging this distinction protects theories that deal in indicators of consciousness from charges of subject changing, as indicators and consciousness itself are two different things. Unlike Integrated Information Theory for instance, which proposes that consciousness *is* integrated information, an indicator approach proposes that consciousness is tied to behavioral indicators, albeit probabilistically. The indicator approach can

only establish probabilistic ascriptions of consciousness, and although this is less than optimal, it is for now the best that we can do, and all that we have ever had.

Necessity and Sufficiency

Dennett's zimbo argument is meant to show that if we present zombie thought experiments honestly, and include all of the behaviors that a true philosopher's zombie would actually engage in, including an "indefinite upward spiral of reflexivity", the zombie would *become* conscious (Dennett 1991, p. 310). This is meant to demonstrate that, in a hypothetical case where we could identify in the behavior of some being *all* the characteristics of a conscious creature, then that creature must be conscious. In Chapter 1, I agreed with Dennett's assertion, so it may seem peculiar that I am arguing that behavior can only provide probabilistic indicators, rather than certain conclusions. I take Dennett's zimbos as being the extreme case in which we have essentially all the behavior that we can have – a perfect clone of a conscious being that is capable of reporting and acting upon its mental states in ways identical to its conscious counterpart. While I think that a hypothetical case like this would get us very close to certainty, cases like these do not come about in the real world. What I'm talking about are situations in the real world, where we need to make inferences based on a limited repertoire of behavior. In the case of Dennett's zimbos, behavior would justify something much closer to a certain judgment, as in the case of a criterial approach. In the actual situation that we find ourselves, all that the behavior justifies is the probabilistic inference that the other is conscious.

To clarify my stance on this issue, I will make explicit my views on what an indicator approach can and cannot accomplish. An indicator approach can establish the

probability of consciousness (PoC) in some system. The way that this is done is through determining which qualities of behavior are necessary, and which are sufficient, to establish a PoC. This is quite different from the question of what qualities are necessary and sufficient *for* consciousness, because the latter discussion would concern constituents rather than indicators. An indicator approach cannot be used to establish with whether a system *is* conscious, but it can be used to establish a PoC. Although this may make the indicator approach appear to be uninteresting and trivial, I do not think that it is. At this point, there is very little consensus and much confusion about how we ought even to begin to justify our intuitions about the consciousness of systems other than ourselves. In this sense the indicator approach provides us with an essential first step – it opens the door to ascriptions of consciousness, and provides support for probabilistic claims.

Since it is possible for a non-conscious system to generate behaviors that appear to be conscious, it will not be possible to make an inference directly from behavior to a PoC. As a result, if we hope to make use of externally observable behavior, I believe that we require a multi-step process. So far, I have argued that mental representation is partly constitutive of consciousness –mental representation is a necessary condition for consciousness. This means that if we are able to identify mental representation as the cause of the behavior of a system, we will be taking the first step in establishing whether the system has a PoC. Since there can be both conscious and unconscious representations, however, it is clearly not sufficient – identifying mental representation-facilitated behavior only gets us half of the way to making an ascription of consciousness. In order to make the inference from mental representation to consciousness, we require an indicator that supports the belief that conscious representation, rather than unconscious representation, underlies the behavior in question. I have some thoughts about what such a behavior might look like,

but I will first evaluate what I take to be the most commonly appealed to indicator of consciousness – namely, verbal reporting.

3.2 - Verbal Reporting

Verbal reporting is taken by some to be the best index of consciousness (Carruthers 1996, Baars 1998, Stoerig 2007). Barring Chalmersian skepticism based on the possibility of the existence of non-conscious clones of conscious creatures, verbal reports do indeed provide some immediately compelling support for the belief that whoever you are conversing with is likely to be conscious. This is because, unlike non-verbal behaviors, verbal reports are perceived as providing detailed information about the content of one's mental states. When my friend says "I feel sad", I can immediately and justifiably come to the conclusion that he is *experiencing* sadness. However, if instead my friend is silent and decides to mope around his apartment, there are a variety of alternative conclusions I could come to – perhaps he is sad, but he may also be frustrated, tired, focused, or maybe he is psychologically atypical and this is how he expresses an emotion such as joy, or excitement. It is thought that non-verbal behavior can have a wide variety of causes and intended effects, whereas verbal reports (whether or not they are truthful) are thought to uniquely reflect the existence of the ability to introspect, and reasonably accurately, or may simply arise out of the having of some experience.

The trust that we place in verbal reports has made the study of animal consciousness a difficult one. Even though most humans have no problem reporting, there is still significant debate about the nature and knowledge of human consciousness, and with

animals that cannot report, the debate is that much more intense. In the following discussion, I will call into question the status of verbal reports as being particularly strong indicators of consciousness, and will propose an approach to evaluating consciousness that takes into account the multitude of ways that consciousness may be expressed.

The Case Against Reportability

Although reporting is considered by some to be the best index of consciousness (Baars 1998 presents it as such), I think that we have good reason to be skeptical of it as a uniquely conclusive indicator. My primary concern has to do with the fact that treating verbal reporting as the best index entails that non-human animals are not conscious. In almost all cases and with respect to almost all species, non-human animals are incapable of reporting on their mental states. As a result, to believe that verbal reporting provides uniquely and supremely valuable insight into whether some being is conscious is to believe that we can never have a well-justified belief that a non-human animal is conscious. Beyond the obvious issue that this idea goes against one of the fundamental goals of this project, it seems highly counterintuitive – why should we feel certain about the conclusion that a sign-language-using gorilla is conscious, and uncertain about the conclusion that its non-verbal mate is not?

It is not that I believe the non-verbal gorilla to certainly be conscious – on the contrary, I think that both cases have a similar degree of uncertainty. This is partly because reporting is a behavior like any other. Whether I point at my dinner and proceed to mime the act of eating, or say to my dinner partner "Let's eat!", I am still engaging in

some behavior. I would certainly concede that using my words constitutes a far clearer method of information transmission, but it is not as if communication of mental states is limited to the domain of verbal reports.

It is also not the case that verbal reporting is a behavior that only conscious creatures can engage in. The voice assistant that is built into my phone is able to communicate verbally, and when prompted, can even make comments on its "mental state" (if I ask the voice assistant how it is doing, it will invariably respond with some version of "very well"). It is more likely than not that my phone is not consciousness, and yet it seems perfectly capable of reporting on its mental states, even if we all agree that it probably has no mental states at all. As AI advances and humans creep closer and closer to building a system that passes the Turing Test with flying colours, this kind of confusing interaction will start to happen more often. A suitably complex non-conscious AI system could conceivably enter into a conversation with a philosopher of mind, and if that philosopher holds the view that verbal reports are on their own strong indicators of experience, s/he is likely to be misled into believing that the AI system is in fact conscious and reporting on real mental states (see Section 3.4 for further discussion of AI)

When comparing verbal reporting to other forms of behavior, I think we have good reason to apply the same level of skepticism across the board. In cases where we know that a system is conscious, verbal reporting can be of more value than other behaviors because it provides uniquely informative descriptions of the contents of mental states. When interpreting verbal reports, it is either the case that the verbal report accurately reflects the reporter's mental state, or that for some reason (deception, ignorance, inaccuracy, etc.) the report is not accurate. When interpreting non-verbal behavior, however, there is no possibility of similar accuracy, because there is no universal standard for the

interpretation of such behavior (other than the flawed notions of folk psychology). Moping, according to some conventions, indicates sadness, but it is not always the case that moping indicates sadness, as it is not necessarily shared cross culturally in the same way that the meanings of English-language words are. It could indicate sadness, but it could also indicate anxiety, uneasiness, bad posture, or nothing at all. Conversely, verbal reports have the potential to provide accurate descriptions of the content of mental states. However, if the goal is instead to determine whether a creature has mental states at all, verbal reports tell us exactly as much as moping, head shaking, or food scavenging does. Verbal reporting is only of special value when the contents, not the existence, of mental states are in question.

This being said, verbal reporting can nonetheless provide clues about whether a system is conscious. When applied to humans (creatures that can already reasonably be expected to be conscious), verbal reports can give insight into the content of their mental representations, as well as provide some support for the conclusion that they are capable of having mental representations. Alternative forms of reporting, such as sign language, can also provide some very compelling evidence when applied to non-human animals. Koko, the remarkably intelligent gorilla who's sign language vocabulary included over 1000 words at the time of her death, was able to communicate with her keepers information such as what kind of food she wanted, what she wanted to do, and even more conceptual things such as the distinction between things that are dead and alive (Morin 2015). In a case like this reporting can be of value, because it happens in the presence of other features and behaviors that are also indicative of mental representation (in the case of gorillas, such features might include their capacity for emotion, or problem solving).

If we are unable to connect verbal reporting, a capacity taken by some to be the gold standard index of consciousness, exclusively to conscious systems, then it seems that we have reason to doubt that any behavior whatsoever can provide us with conclusive information about whether some creature is conscious. The issue here is that in all cases but our own, behavior is *all* that we have. We are therefore left with two options – either we accept that ascriptions of consciousness are impossible unless applied to oneself, or we formulate a behavioral approach that can mitigate the problem of uncertainty. Certainly, before accepting the rather unappealing hypothesis of option 1, we ought to give option 2 our best shot.

What might such an approach look like, though? To start, we should take inventory of what we know to be true of consciousness. One thing that is obvious about consciousness is that it can manifest itself in a wide variety of ways – the ways that we (conscious humans) act is wildly variable. Furthermore, as far as we know, there is no behavior that is unique to conscious creatures, or necessary for conscious creatures to engage in. What we have here with consciousness, then, is a concept (or category) that seems to have no defining characteristic behavior. Consciousness is certainly not the only concept that has this character, and luckily for us, Wittgenstein has helpfully provided the philosophical community with tools to address these kinds of issues.

3.3 - Categorization and Soft Constraints

Ascriptions of consciousness are categorizations. When I claim that my friend Joe is conscious, I am placing him into the category of "conscious systems". I am doing so

because Joe has displayed enough attributes associated with that category to justify my placing him there. But on what basis do I determine what these attributes are, and that Joe has met enough of them to be considered a member of said category?

These questions relate to some general questions about categorization. In some cases, we are unable to identify a single attribute shared by all members of a category. Take the category of "can openers", for instance – it might be proposed that a necessary and sufficient condition that something must meet to be considered part of this category might be the condition of "being a tool that may be used to open cans". If there could ever be a defining characteristic of can openers, this would probably have to be it. If something meets this single condition, it ought to be considered part of the category no matter how unique or unlike the other members it might be. The problem is that, despite not being a can opener, a sledgehammer is very much capable of opening cans, even if it does so in a messier way than other things that can open cans. Furthermore, it seems wrong to consider a broken can opener (something that may no longer be used to open cans) to *not* be a can opener. A can opener doesn't stop being a can opener when it breaks, it just becomes a dysfunctional can opener.

If a concept as simple as can openers can be this convoluted, what hope do we have of understanding a concept as complex as expressions of consciousness? As I have discussed in Chapter 2, I believe that the most we can say with certainty about consciousness is that it has a representational base. There may be many other features to consciousness that need to be present in greater or lesser degrees depending on the system, but minimally, representation is the common thread. The category of *expressions of consciousness* does not seem to have even this measure of continuity, because there does not seem to be any single feature that can be found in every instance of behavior that is an expression of

consciousness. In cases like these, a plausible alternative to the kind of straight-forward categorization mentioned above is that of the family resemblance, or soft constraint.

Some of the most influential early work done on the topic of cluster concepts is that of the later-Wittgenstein, and his much-referenced investigation of the concept of "games" (Wittgenstein 1953, Remark 66). The question of what makes some activity a "game" is a deceivingly difficult one to answer. The first common characteristic that is often brought up is the idea of enjoyability – it is intuitive to think that games share the common trait of being enjoyable. But this isn't always the case. If a group of people who collectively despise board games voluntarily decide to sit down and play one, despite receiving no enjoyment from the activity, most would agree that they are nonetheless playing a game. Perhaps one could turn to the notion of "rules" and claim that all games share the characteristic of being activities that involve a set of rules that must be followed to facilitate the game. This criterion does not neatly map onto imagination-heavy games played by children – such as in the case of "house", or "doctors and nurses" – where the rules, if there are any, are poorly defined and open to change. Furthermore, there are many activities that, despite being enjoyable or involving rules, are not considered games (watching your favorite television show, speaking in a grammatically correct way, etc.).

This roadblock to simple categorization led Wittgenstein to propose the notion of family resemblance,

> I can think of no better expression to characterize these similarities than "family resemblances"; for the various resemblances between members of a family; build, features, colour of eyes, gait, temperament, and so on and so forth overlap and criss-cross in the same way. (Remark 66)

Here Wittgenstein is comparing the concept of games with that of families, in the sense that members of the same family may share some features and not share others. It is not as if to be considered part of a family one is required to have the same nose, or the same dispositions, as other members of the family. Instead of sharing one core feature, on Wittgenstein's model members of a family may be related by a network of shared features. Wittgenstein uses the helpful analogy of a steel bridge cable: although the cable may be perceived as being one continuous piece of material, it is in fact made up of a series of smaller, interwoven wires – none of which extend the full length of the cable itself. In the same way, shared features between members of a category form complex networks, with the potential for there to be no single feature that is shared by all members. This discussion of categories is making use of the idea of the "cluster concept" – a concept that is defined by numerous attributes, such that the satisfaction of no single attribute is necessary or sufficient for membership.

The notion of soft constraints relates closely to the family resemblance theory, though it has typically played a more prominent role in cognitive science than it has in philosophy (Dawson 1998, p. 38,51, Thagard 1996, p. 126). Applied to articulating concepts, it is the notion that a concept will be the name for something in virtue of an open-ended set of properties – if the thing has enough of these properties, it will be an example of what the concept names. The reason that a constraint would be considered soft, rather than hard, is that they can be violated without necessarily disqualifying something from membership in a category. An example of a soft constraint might be the anti-bacterial quality of soap – it is not necessary for a substance to be considered "soap" that it be anti-bacterial. On the other hand, a hard constraint (an inviolable, necessary condition) on the concept of soap might be that it functions as a cleaning agent. If a company decided to

market a new brand of soap that they advertised as "not used for cleaning", then this would be a violation of a hard constraint. The indicator approach deals in soft constraints rather than hard ones, and this allows for it to be easily applied to a variety of different species with vastly different behavioral patterns.

A continuation of Wittgenstein's work on family resemblance can be found in Eleanor Rosch's *Prototype Theory* (Rosch 1973). Rosch suggests that categories are based on a set of attributes that are commonly (but not universally) shared amongst members of said categories, and that certain attributes are more central to membership in said categories than others. For example, a common attribute shared amongst most, but not all, of the avian species is the ability to fly, and as a result, it is a central and heavily weighted attribute when considering what to include in the category of "birds". Weighing less than one pound is also an attribute that is found amongst members of the category, but it is not as widespread, and therefore not as central or relevant to membership as being able to fly. When looking for indicators to establish a PoC, Rosch's work can provide a helpful guideline for interpreting which behaviors are relevant, and which are not.

When trying to understand the mental states of others, we often arrive at what we believe are well-supported conclusions on the basis of behavioral evidence. An example of this can be found in a recent study on animal consciousness, where the authors provide the following justification for the inherent uncertainty in their conclusions,

> Rather than emphasizing a particular indicator as being decisive, we propose that the consistency amongst these indicators can serve to assess consciousness in particular species. The integration of scores on the various indicators yields an

> overall, graded criterion for consciousness, somewhat comparable to the Glasgow Coma Scale for unresponsive patients. (Pennartz et al 2019, p.1)

The authors point out that approaches that deal in indicators, or soft constraints, are not without precedent in the field of medicine, where doctors must assess the responsiveness, or consciousness, of patients. The Glasgow Coma Scale operates in much the same way – where the responsiveness of a patient is evaluated across a variety of different capacities as an indirect measure of the consciousness of an individual. This kind of behavioral, symptom-based approach is a perfect example of soft constraints at work, since in most cases, a diagnosis can be arrived at confidently even when some symptoms are missing. Although I would resist the authors use of the term "criterion", I am certainly sympathetic to the idea that evaluations of consciousness ought to be graded, in the sense that more instances of mental representation and behavioral flexibility increase the PoC.

To be clear, I do not mean to propose that consciousness *itself* is a family resemblance, or cluster, concept. It may turn out to be, but I have argued thus far is that we have good reason to believe that mental representation is a common feature across all systems capable of consciousness, thereby preventing it from being considered at least entirely a cluster concept. *Expressions* of consciousness, on the other hand, are an entirely different matter, because there is no one characteristic that is common to all behavior that a conscious creature engages in.

While the fact that there is no one common behavior that is criterial to all consciousness does pose a problem, it is a problem that can be partially overcome if we can identify a characteristic in behavior that is reliably associated with consciousness. I believe that behavioral flexibility may be just such a characteristic.

3.4 - Behavioral Flexibility

An ability that shows significant promise as an indicator of a probability of conscious representation is that of *behavioral flexibility* (also referred to as flexible control). Behavioral flexibility refers to the ability of a system to respond to novel circumstances, in ways that are responsive in detail to changing stimuli, and not dictated by instinct or by genes. Behavioral flexibility allows for the development of new behaviors as well as the modification of previously learned behaviors, and is considered by some to be the "hallmark" of consciousness (Price & Norman 2008, p.30).

Flexibility is usually contrasted with the more rigid class of automatic behavior, which is believed to be less responsive to environmental variables, novel situations, and volition. Automatic behavior is commonly instinctual and perhaps only minimally adaptable to unfamiliar circumstances. As such, flexible control is quite useful when there is no suitable instinctual behavior for the task at hand – a capacity that gives a significant evolutionary advantage. It has been proposed that consciousness may have actually evolved in order to facilitate flexible control, as the ability to respond in a non-automatic way to environmental stimuli may be quite advantageous to one's survival. In a discussion of the biological function of consciousness, Brian Earl writes,

> Consciousness is associated with a *flexible response mechanism* (FRM) for decision-making, planning, and generally responding in nonautomatic ways. The FRM generates responses by manipulating information and, to function effectively, its

> data input must be restricted to task-relevant information. The properties of consciousness correspond to the various input requirements of the FRM; and when important information is missing from consciousness, functions of the FRM are adversely affected… (Earl 2014, p.1)

Here Earl highlights the two-way relationship between consciousness and behavioral flexibility – where properties of consciousness correspond to properties of behavioral flexibility, and vice versa. Earl suggests that the FRM developed out of an evolutionary requirement to be able to respond to unique circumstances, where unsuitable automatic behavior had the potential to put the life of the organism at risk or limit its ability to reproduce. He goes on to propose that the role of consciousness is to facilitate the FRM by synthesizing relevant environmental data and feeding this data into the FRM.

Similar claims have been made about *volitional movement* – an ability that is sometimes used as an example of flexible control. When we discuss volitional movement, what we are referring to is the ability to make decisions about the way that we move and behave based on our beliefs and desires. In their investigation into the biological function of consciousness, Lee M. Pierson and Monroe Trout echo claims made by Earl in their conclusion that consciousness facilitates volition:

> [I]f consciousness has any biologically adaptive function, it must ultimately be for some type of movement. On the one hand, there are good theoretical arguments and empirical evidence that volitional movement requires consciousness. On the other hand, we know of no good theoretical arguments or empirical evidence that automatic movement requires consciousness. There is also empirical evidence that

> conscious experience varies concomitantly with volition. Volitional movement is adaptive because it adds a fundamental flexibility that is not possible with deterministic, automatic movements, which are controlled entirely by neural processes. (Pierson & Trout 2017, p.70)

Here Pierson and Trout argue that empirical evidence supports the claim that volitional movement cannot happen in the absence of consciousness. I think that this claim has some intuitive appeal – it is not immediately clear what volition might look like in a non-conscious system. We can understand what *unconscious* volition might look like (unconscious in the Freudian sense of the term, as it relates to unacknowledged beliefs or desires that nonetheless have an impact on one's behavior), but only insofar as it would have a causal effect on one's conscious states. The authors also make the claim that:

> All programs, even neural net or other "learning" programs, lack the flexibility that volition can provide. Although such programs can be written to modify themselves in accordance with certain new conditions–to "learn" to adapt– they can adapt only to new conditions for which they are programmed. Adding self-modification algorithms to the original program merely pushes the problem up to a higher level. The modifications that can be made by a self-modifying program are always constrained by its original algorithm. (p.65)

If the authors are right, then any behavior that demonstrates volitional control must indicate consciousness, though I believe that AI may give us reason to doubt the strength of their conclusion (see below for further discussion of this issue).

Whether or not these theories are correct in their conclusions about the specific biological roles of consciousness and behavioral flexibility, it is clear that the two theories are quite closely linked in one way or another. One of the characteristics of a conscious experience, at least in the case of humans, is the ability to control attention and engage in active control over behavior. When I am conscious of some element of my environment, I am able to factor that element of my environment into my conscious decision-making process and make decisions on the basis of that element. Furthermore, I am able to flexibly control my response, in the sense that I am not required to act in any particular way should I have reason not to. Deciding to get fast food instead of a healthy meal may not be a wise choice, but it is a choice that I can make, along with many others depending on my mood, emotional state, and rational evaluation of my circumstances. Behavioral flexibility is necessary for volitional control over action, and a lot of our encounters with our own consciousness seem to happen in the context of volitional control. Behavioral flexibility is also an ability that is objectively observable and detectable, which makes it particularly useful when it is used in our approach for determining behavioral indicators of the PoC. One common way that flexibility can be established and measured in a non-human animal is through a "reversal learning paradigm", where animals are taught a specific kind of novel behavior that is rewarded with food, and then required to un-learn said behavior in order to continue receiving their rewards (Audet & Lefebvre 2017. P.941). Reversal learning tests have been given to rats, where the rats are taught to stick their noses into specific holes in order to receive food (Bari et al 2010). Once the rats have memorized this behavior, the food producing holes are switched around, and the rats have to re-learn which holes produce food and which do not. Tests of this sort establish flexibility by determining whether the test subject is able to take in changing environmental stimuli

and respond in a non-automatic way. If the rats were not able to flexibly control their behavior, then they would presumably be unable to correct the behavior that they have developed that no longer results in the successful acquisition of food. Behavioral flexibility requires a cycle of active and volitional re-evaluation not possible in automatized behavior, where new patterns can be recognized, and novel problems can be understood. This capacity, when paired with the more basic capacity for mental representation, provides the kind of cognitive environment that is highly indicative of consciousness.

I should make explicit that I do not mean to claim that all conscious behavior *must* be flexibly controlled. There may be fully conscious behavior that can happen automatically or inflexibly within a system. Octopuses, for instance, are thought to be conscious of the movements of their tentacles without having fine grained control over them. While an octopus can guide its tentacles in a general direction, much of the nuance of its movement happens as a result of the rather independent local nervous systems located in each tentacle (Godfrey-Smith 2017, p.71-72). We can also be very much conscious of our heartbeat, without having flexible control over the rate at which our hearts beat (though it is possible for this ability to be learned). Rather, behavioral flexibility can act as a plausible indicator of a particular, limited kind of consciousness. The presence of flexible control implies a system that evaluates and re-evaluates its behavior in response to changing interoceptive representations, beliefs and desires, and environmental stimuli – an active and engaged process that is indicative of conscious mental representation and even difficult to make sense of without an appeal to conscious volition.

Importantly, flexible control is not an all-or-nothing concept. There are degrees of flexible control, in the sense that we may have greater volitional control over some behaviors than others. It may very well be the case that consciousness is similarly scalar, and

that the degree of an organism's conscious experience could be affected by the degree of flexible control that is has, as well as the kinds of perceptual information that the organism has access to. If this is the case, then an approach such as the one that I am describing could be used to gauge not just the probability that some organism is conscious, but also *to what extent* an organism is conscious. Determining how the approach could be applied to function in this way is beyond the scope of this project, but I do believe that a commitment to the indicating power of flexible control entails a commitment to the view that consciousness is scalar – a view that seems intuitively plausible and that I am quite sympathetic to.

The AI Challenge to Flexible Control as an Indicator of Consciousness

Although flexible control can certainly be seen as a behavior that is highly indicative of a high PoC, it faces at least one serious challenge – the zombie problem. In recent years, artificial intelligence has made some major progress, and there are now consumer-level A.I. products that appear to demonstrate degrees of flexible control. Driverless vehicles are an extreme example of this idea – they are able, through a variety of sensors and input streams, to synthesize information about their rapidly changing external environment as well as their "internal" states (tire pressure, fuel consumption, etc.), and adjust their behavior to accommodate for these changes in a flexible way. This is of course necessary, because road conditions are constantly changing and rarely predictable. To stretch this idea even further, abilities that have traditionally been associated with conscious creatures, such as goal directed behavior or the ability to learn, are very much

present in these driverless vehicles. This is an issue for our use of flexible control as an indicator of conscious representation, because if non-conscious creatures are capable of flexible control, then it may not be the case that flexible control is a sufficient condition for the the PoC.

I do think there are some ways around this issue. One response might be that the premise of the objection is false – and AI system is not actually engaging in flexible control, just something that *looks* like flexible control (this is the position of Pierson & Trout mentioned in the previous section). It may appear to us that driverless cars are able to respond to changes in their environment in a way that is truly flexible, but this might not actually be the case. For instance, I have the choice at any time while driving my car to veer into oncoming traffic. I am not likely to make this choice, but it is a choice that I can make. A driverless car on the other hand does not have this capacity. It is capable of choosing from a pre-programmed repertoire of movements and would not be capable under any circumstances of making the decision to drive into oncoming traffic (and rightfully so). As a result, interpreting what the car is doing as anything but automatic and non-volitional may be flawed from the very start. This line of argumentation might preserve flexible control for the time being but would require a secondary argument against the prospect of a non-conscious AI system ever having what we would consider to be "truly" flexible control. One would of course also have to argue for the claim that humans *can* have truly flexible control (a hotly debated topic), which would have to be accompanied by an argument as to whether such an ability is limited to organic creatures (also a hotly debated topic). I do not find arguments of this sort to be compelling, as I have seen no evidence to suggest that organisms have the unique capacity for flexible control. If an AI system can demonstrate what *looks* like flexible control, to the extent that we would not

hesitate to ascribe flexible control to an organism engaging in similar behavior, then we ought to treat the AI system as though it is engaging in the same kind and quality of behavioral flexibility as the organism.

An argument could be made that any approach that we come up with to evaluate the consciousness of some creature is going to produce inaccurate results when applied to AI systems. The intention behind AI is for AI systems to appear as human, or in other words as conscious, as possible, and this complicates matters greatly. If the point of some AI systems is to make us believe that they are conscious when they are not, then this obviously makes them difficult to evaluate. In this sense the motivation behind the creation of AI is rooted in a kind of deception – an attempt to come as close to passing the Turing test as possible – and this could be done equally well with a very robust non-conscious AI system as it could be with a truly conscious AI system. This is not the case for humans and non-human animals, as consciousness is likely to play a functional biological role in such creatures that makes possible the kinds of behaviors that AI systems may exhibit without the need for consciousness.

I do not find arguments of this sort to be very compelling. If we are to adopt any approach to evaluating whether some system is likely to be conscious, it really ought to apply evenly to all systems that could be conscious. If one would want to claim that the indicator approach is not compatible with AI, one would either need to argue that AI could never be conscious (a claim that I believe is unlikely to be true given the rapidly increasing sophistication of AI technologies), or argue that the nature of AI means that we can never know whether AI systems are conscious (a claim that would need some defense, and that does not immediately appear plausible to me). It is true that we may be able to find examples of AI systems that we believe to be non-conscious that also engage in behaviors

that indicate mental representation and behavioral flexibility, however such examples are likely quite rare. This is not a problem for my approach because the intention behind the approach is simply to establish the PoC, not the certainty of consciousness. We may be able to identify other things about the AI system that give us reason to believe that the PoC is low, such as limited flexible control in other behavioral domains. If we determine that some AI system has demonstrated behavior that is indicative of mental representation and flexible control, but we nonetheless have good reason to believe that the AI could not be conscious, then these reasons would simply result in a low PoC. Conversely, we may raise the PoC based on other considerations, such as the sophistication of the hardware that the AI system relies upon, and the ability of the AI system to demonstrate behavioral flexibility across multiple domains. The same can be said of applying the approach to non-human animals – we may be able to establish that jellyfish have a PoC, but the PoC can be lowered based on other considerations, such as the fact that jellyfish lack a centralized nervous system. Until such a time as it is proven that AI systems will *never* become conscious, we should not hesitate to apply the same approaches to artificial intelligence as we do organic intelligence, despite the potentially uncomfortable nature of our conclusions.

3.5 - Application of the Approach to Tool Use

We are now ready to begin thinking about what kinds of behaviors provide justification for the belief that some system has a high PoC. In Chapter 2, I argued that mental representation is a necessary condition for consciousness, though it remains unclear

exactly what features differentiate the conscious representations from the unconscious. In this chapter, I have argued that behavioral flexibility is a feature of externally observable behavior that can be used to determine the probability that the representations that underlie that behavior are conscious. If I am correct, then we can construct the following approach to evaluating behavior:

Externally observable behavior suggests a PoC when it:

1. Requires the capacity for mental representation *and*
2. Demonstrates the capacity for flexible control.

In the first step, we need to draw a connection between observable behavior and mental representation, by positing mental representation as necessary to explain the behavior. In the second step, we need to identify behavioral flexibility in order to determine whether the representations that underlies the behavior are likely to be conscious. With this in mind, I will now demonstrate though the example of tool use how this approach can be applied in a very specific context.

Tool use may be the archetypal example of a behavior that meets both of our conditions, as it is one specific kind of flexible behavior. Take the following example of a crow using elements of its environment in a novel way to reach food, for instance:

> New Caledonian crows were presented with Bird and Emery's (2009a) Aesop's fable paradigm, which requires stones to be dropped into a water-filled tube to bring floating food within reach. The crows did not spontaneously use stones as tools, but

> quickly learned to do so, and to choose objects and materials with functional properties. Some crows discarded both inefficient and non-functional objects before observing their effects on the water level (Taylor et al. 2011, p.1)

Consider the most general kind of mental representation required of the crows in this case. The crows are taking in visual and tactile information from their environment and synthesizing it in such a way that they can then evaluate the state of their environment with respect to their goal, which is acquiring food. This is a relatively low-level cognitive ability, and the crows likely share this ability with many other lower-level organisms (single celled organisms operate in a similar way, detecting chemicals in their environment and moving further away from harmful chemicals and closer to chemicals that sustain them). What is interesting and indicative about the behavior of the crows is that they are able to understand the abstract cause and effect relationship between their actions and the proximity of the food. This cause and effect relationship is indirect – the crow is engaging with elements of its environment that are not directly related to the food that it is trying to reach. The ability of the crow to interact with novel items and form beliefs about their level of usefulness seems to indicate a level of understanding that goes beyond automatic movements. Furthermore, the environmental resources that the crows are making use of are new to them, meaning that they were all able, in a very short period of time, to learn about this novel and abstract cause and effect relationship and make attempts to engage with it, even going so far as to not bother with elements of their environment that did not have the relevant water-displacing property.

Another experiment on this same breed of crow was conducted to test their ability to understand problems that were not within their field of view. The authors summarize their findings in the following passage, which I have abridged in the interest of brevity:

> Crows were able to mentally represent the sub-goals and goals of metatool [multiple tool requiring] problems: crows kept in mind the location and identities of out-of-sight tools and apparatuses while planning and performing a sequence of tool behaviors.
>
> …
>
> Our results provide conclusive evidence that some NC crows can preplan using mental representations of the sub-goals and goals of a metatool problem … In experiment 1, we found that most of the crows we tested solved a stick problem, where they had to use a stick tool to get a stone tool and then use the stone to get food while avoiding a distractor object (another stick). This was despite crows not being able to view more than one stage of the problem at a time. Four of the crows we tested showed clear evidence of preplanning. (Gruber et al. 2019, p.689)

Although the results of the authors' crow study are quite interesting, they do not provide much justification for why they believe that the behavior of the crows is indicative of mental representation. They assert that the crows must have used representation to pre-plan behavior and solve multi-stage problems, but I could find no arguments to motivate these strong claims. This being said, I do believe that the authors have identified behavior that is indicative of representation and even difficult to explain without it, so I will attempt to provide some justification for this claim here.

In order to pre-plan their behavior, the crows must be making use of some sort of memory – the only way for them to follow their pre-made plan would be for them to recall what that plan was. It is difficult to imagine what memory would look like without mental representation. When we recall a memory, we are generating a mental representation of a past event. What exactly the role is of these representations is debated, but there is general consensus about the representational nature of memory. If memories are indeed representational, then any behavior that displays a utilization of memory can fulfill the requirement of our first step. The ability to keep in mind locations of out-of-sight objects also indicates mental representation, as this is plausibly an ability that requires the capacity to remember locations and objects. Furthermore, for those acts of memory recollection that involve volition, such as in the case of memories that are recalled on demand as needed to address some problem, such behavior could also justify the inference to conscious representation, which is our second step, since volitional control over action indicates a PoC.

Pre-planning also requires the ability to plan into the future, and to do this effectively, the crows would have to be able to anticipate the cause and effect relationships between themselves and elements of their environment. This indicates the ability to visualize and anticipate different potential actions and outcomes – a behavior that is also difficult to make sense of without an appeal to mental representation.

With mental representation having been identified as the cause of the behavior of the crows, we have made the first step towards determining whether there is a PoC. The next step, which is to connect representation to consciousness, requires us to identify flexibility in their behavior, as conscious behavior commonly includes movements that are under the flexible control of the system. I think that we have good reason to believe that the

crows are in fact displaying behavioral flexibility. In both cases, the test subjects appear to demonstrate the capacity to override automatic movements based on novel information, an ability that is closely linked to the reversal learning tests previously discussed. In the first case, the crows make selections based on the efficacy of the tools at displacing water, a process that would have to be informed through the crow engaging with novel elements of its environment and coming to some sort of decision. Without consciousness, this degree of behavioral flexibility is very difficult to make sense of.

In the second experiment we seem to have an even stronger case for flexibility, as the crows were able to factor novel information about the tools that they were provided into their plans. The authors note that, although tools that the crows might have evolved to use (such as sticks) were provided, the crows nonetheless factored the unfamiliar experimenter-provided tools into their decision making when these tools were better suited for the task at hand (p. 690). This behavior points to an ability to flexibly control behavior based on the contents of the crows' mental representations. Furthermore, the multi-step nature of the problem presented to the crows, and the trial and error method through which they were reliably able to solve it, indicates that the crows have the capacity to flexibly control behavior in response to an unfamiliar environment. In other words, the crows seem to be demonstrating real understanding of their circumstances, rather than automatic behavior.

These experiments provide the insight needed to complete the two-step process that I have proposed. We can complete the first step because the crows demonstrate behavior that is difficult to make sense of without an appeal to mental representation, as the crows are capable of keeping in mind locations and functions of tools that were not within their line of sight. We can also complete the second step, because the ability of the crows

to respond to novel environmental elements is highly suggestive of conscious mental representation. In this sense, the type of tool use that the crows engage in meets the necessary and sufficient conditions that I have laid for establishing a PoC. This PoC could be modified based on the results of future experiments that provide insight into the ability of the crows to use mental representation and flexible control in other behavioral domains.

3.6 - Conclusion

Although verbal reporting is often held up as the gold standard indicator of the PoC, we have good reason to doubt its adequacy when it is used as the sole means of evaluation. One of the biggest problems with reportability is that it rules out the possibility of non-human animal consciousness. The capacity for mental representation and behavioral flexibility in an organism provides strong support for a PoC. Representation is a necessary condition of consciousness, and as such, the identification of mental representation as the cause of behavior of some system serves as an important first step to establish the PoC. In turn, behavioral flexibility can give us insight into whether the representation underlying some behavior is likely to be conscious, as we have good reason to believe that producing flexibility in behavior is a common characteristic of a particular kind of conscious representation. The indicator approach can therefore help to establish the PoC in systems other than ourselves. This approach provides a practical and flexible framework for the interpretation and evaluation of externally observable behavior, particularly the kinds of non-verbal behavior that non-human animals are limited to. Tool use is a behavior that is highly indicative of consciousness, because it is an activity that gives us evidence of mental

representation as well as behavioral flexibility, and indeed, it is difficult to explain such a behavior without an appeal to both of these consciousness-indicating faculties.

Conclusion

Determining whether some creature other than oneself is conscious first involves a commitment to the kind of thing that consciousness is. I have given three main arguments against anti-cognitivist accounts of consciousness. In the first, anti-cognitivist accounts of consciousness are at odds with Occam's Razor – an account that takes consciousness to be a non-cognitive phenomenon is equally unsupported if it includes one non-cognitive substance as it is if it included a million non-cognitive substances. In the second, Dennett's version of a fully realized phenomenal zombie, the "zimbo", provides a strong argument against the metaphysical possibility of a zombie. Zombie arguments, such as those of Chalmers, are frequently used to argue for anti-cognitivist accounts, and Dennett's approach provides good reason to doubt such arguments. In the third, particular versions of the mind-body and mind-mind problems that anti-cognitivism creates for itself deal what I take to be the final blow to anti-cognitivist theories. We have no framework to understand what cause and effect might look like in the absence of spatiotemporal relationships between objects (Kim 2001), and as such, no way of conceptualizing non-cognitive causation. I have found no plausible explanations of what non-cognitive causation might look like, indeed the best that an anti-cognitivist account can provide us with is a version of epiphenomenalism – a position that I take to be non-starter. I argue that these issues give us

good reason to doubt anti-cognitivist approaches and, in the spirit of Occam's Razor, take cognitivist theories to be the most reasonable position.

Starting with the conclusion that consciousness is a cognitive phenomenon, in the second chapter I begin to explore what cognitive features might characterize consciousness. I suggest that approaches to articulating what it means to be conscious typically fail in two ways, and illustrate these common routes to failure as two horns of a dilemma. The first horn (circularity) concerns definitions of consciousness (such as Tye's "simple claim"), in that most of the time they are so narrowly circular that they cannot tell us anything interesting about what consciousness is. The second horn (subject changing) concerns explanations of consciousness, where cognitive accounts of consciousness often describe cognitive functions that could quite easily happen in the absence of experience, rather than the functions that underlie experience. The result of this misstep is a changing of subject from consciousness to something else. I have not seen this dilemma articulated in these terms elsewhere, and I take my explanation of the two horns of the dilemma to be one of the major contributions of this thesis. I suggest that one way to retain the important elements of each horn, while avoiding the problematic ones, is to adopt what I refer to as the MC – that consciousness has a *representational base*. This is not a claim that representation is all that consciousness is (I am neutral on that issue), rather, it is the claim that representation is a base-level necessary condition for consciousness. I take this claim to be uncontroversial and widely agreed, as representation is a common feature to almost all contemporary theories of consciousness, both anti-cognitive and cognitive.

In the third chapter, I suggest that the MC can play a useful role in evaluating whether externally observable behavior indicates a likelihood of consciousness. There are a great many behaviors that require mental representation to be made sense of, such as

those that require memory and pre-planning of action. I argue that identifying mental representation as the cause of some behavior constitutes an important first step towards establishing a probability of consciousness (PoC). However, a second step is required, one that connects representation to consciousness, as there can be both conscious and unconscious representations.

I suggest that one plausible indicator of conscious mental representation is *behavioral flexibility* – a system responding in a non-automatic and non-instinctual way to novel stimuli. Behavioral flexibility allows for the creation of novel approaches to problem solving (such as in the case of the New Caledonian crows discussed in Chapter 3). It is the capacity for evaluation and re-evaluation of behavior in response to changing environmental and interoceptive states – these are indicative of consciousness, and even difficult to make sense of without an appeal to consciousness (particularly so in the case of organisms, as it has been proposed that consciousness actually evolved to facilitate flexible control). Behavioral flexibility also has the advantage of being straightforwardly testable through experiments such as the reversal learning paradigm. As such, I suggest that one method to establish a PoC based on externally observable behavior is through the identification of (1) mental representation and (2) behavioral flexibility.

I conclude by applying this method to one kind of flexible behavior, *tool use,* and argue that, at least in the experimental cases cited, tool use indicates both mental representation (pre-planning of action and maintained awareness of environmental factors that were out of sight), as well as behavioral flexibility (use of novel tools to solve unfamiliar problems through methods that could not be fully accounted for through appeal to automatic, instinctual movement).

I believe that one of the primary contributions that this thesis has to offer the field of consciousness research is that it brings together and organizes a wide variety of diverse and sometimes confusing material in the literature. Wilfrid Sellars, in his book *Philosophy and the Scientific Image of Man*, writes "The aim of philosophy, abstractly formulated, is to understand how things in the broadest possible sense of the term hang together in the broadest possible sense of the term" (Sellars 1962). This project was undertaken very much in the spirit of this idea – that there is value in providing an organizational structure to ideas that otherwise may appear disparate or confusing. Consciousness as a research topic is perhaps even more in need of such treatment than other sub-disciplines in philosophy, since there is such a great degree of overlap and nuanced definitional differences amongst competing theories.

I anticipate this project developing in a number of ways. I alluded to the modulation of the PoC depending on other factors that will either lend support to or undercut an ascription of consciousness, but I do not explore in any detail what these factors might be. It will be important to establish what these factors are and how they apply in order to develop this approach into something that can reflect the scalar nature of consciousness (or perhaps, to determine whether consciousness is scalar in the first place). Discussions of the scalar nature of consciousness also raise important ethical questions, such as the question of how conscious a creature needs to be in order to be morally relevant.

In recent years, as more research has been done on the topic of flexible control, some have argued that there may be good reason to believe that certain "automatic" behaviors do have elements of flexibility. This suggests that further work will need to be done on the topic of behavioral flexibility and non-automatic behavior to determine whether and how to make a distinction between the two kinds of flexibility. I have also

proposed behavioral flexibility as being only *one way* of identifying conscious mental representation via behavior – there is certainly more work to be done in this regard, since it seems possible that there are a great many other behavioral indicators that could be identified.

References

Audet, Jean-Nicholas. Lefebvre, Louis. (2017). What's Flexible in Behavioral Flexibility?, in *Behavioral Ecology.* Vol. 28:4. P. 943-947. Oxford University Press. Oxford, UK.

Baars, Bernard J. (1998). *A Cognitive Theory of Consciousness.* Cambridge University Press. Cambridge, UK.

Baars, Bernard J. (2005). Global workspace theory of consciousness: toward a cognitive neuroscience of human experience?, in *Progress in Brain Research.* Vol. 150. P. 45-53. Elsevier Publishing. Amsterdam, NL.

Block, Ned. (2011). The Higher Order Approach to Consciousness is Defunct. In *Analysis.* Vol. 71:3. P. 419-431. Oxford University Press. Oxford, UK.

Block, Ned. (2002). Concepts of Consciousness. In D. J. Chalmers (Ed.), *Philosophy of Mind: Classical and Contemporary Readings.* P. 207-218. Oxford University Press. New York, USA.

Block, Ned. (1995). On a confusion about a function of consciousness, in *Behavioral and Brain Sciences.* Vol. 18. P. 227-287. Cambridge University Press. Cambridge, USA.

Brook, A., & Raymont, P. (2019). *A Unified Theory of Consciousness.* Unpublished manuscript. Carleton University. Ottawa, CAN.

Brook, A. (2006). Kant: A Unified Representational Base for All Consciousness. In Uriah Kriegel & Kenneth Williford (Eds.) *Self-Representational Approaches to Consciousness*. P. 89-110. MIT Press. Cambridge, USA.

Carruthers, Peter. (1996). *Language, Thought, and Consciousness: An Essay in Philosophical Psychology.* Cambridge University Press. Cambridge, UK.

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

Chalmers, David J. (2002). Consciousness and Its Place in Nature. In David J. Chalmers (Ed.) *Philosophy of Mind: Classical and Contemporary Readings.* P. 247-272. Oxford University Press. New York, USA.

Chalmers, David J. (2013). *Panpsychism and Panprotopsychism*. Amherst Lecture in Philosophy. Amherst College. Amherst, USA.

Clark, Andy. (2001). *Mindware.* Oxford University Press. Oxford, UK.

Dawkins, Marian. (2014). Animal Welfare and the Paradox of Animal Consciousness. In *Advances in the Study of Behavior*. Vol. 47. Elsevier. Amsterdam, NL.

Dawson, Michael R. W. (1998). *Understanding Cognitive Science*. Blackwell Publishers Inc.. Malden, USA.

Delancey, Craig. (2006). Basic Moods. In *Philosophical Psychology.* Vol. 19:4. Routledge Press. London, UK.

Dennett, Daniel C. (1991). *Consciousness Explained.* Back Bay Books/Little, Brown and Company. New York, USA.

Dennett, Daniel C. (1991). *Brainstorms.* The MIT Press. Cambridge, USA.

Dretske, Fred. (1995). *Naturalizing the Mind.* MIT Press. Cambridge, USA.

Earl, Brian. (2014). The Biological Function of Consciousness. In *Frontiers in Psychology.* Vol. 5: 679. Frontiers Media. Lausanne, CH.

Frässle, S., Sommer, J., Jansen, A., Naber, M., and Einhauser, W. (2014). Binocular rivalry: frontal activity relates to introspection and action but not to perception. In *J. Neurosci*. Vol. 34. P. 1738–1747.

Frege, Gottlob. (1892). *Uber Sinn und Bedeutung*. Trans. By Max Black. P. 25 – 50.

Gallagher, S., and Zahavi, D. (2007). *The Phenomenological Mind: An Introduction To Philosophy of Mind and Cognitive Science*. Routledge. New York, USA.

Godfrey-Smith, Peter. (2017). *Other Minds: The Octopus, The Sea, and The Deep Origins of Consciousness*. Farrar, Straus, and Giroux. New York, USA.

Goldie, Peter. (2000). *The Emotions: A Philosophical Exploration.* Oxford University Press: Oxford, UK.

Gruber et al. (2019). New Caledonian Crows Use Mental Representations to Solve Metatool Problems. In *Current Biology.* Vol. 29. P. 686-692. Elsevier. Amsterdam, NL.

Harman, Gilbert. (1990). The Intrinsic Quality of Experience. In J. Tomberlin (ed.), *Philosophical Perspectives.* Vol. 4. P. 31-52. Ridgeview Publishing Company. Atascadero, USA.

Honderich, T. (2006). Radical Externalism. In A. Freeman (Ed.), *Radical externalism: Honderich's theory of consciousness discussed*. P. 3–13. Imprint Academic.

Huxley, Thomas (1866). *Lessons in Elementary Physiology*. Macmillan. London, UK.

Jackendoff, R. (1987). Consciousness and the Computational Mind. In *Explorations in Cognitive Science.* No.3. The MIT Press. Cambridge, USA.

Jackson, Frank. (1982). Epiphenomenal Qualia. In *The Philosophical Quarterly*. Vol. 32:7. P. 127-136. University of St. Andrews. St. Andrews, SCT.

Jackson, Frank. (2007). The Knowledge Argument, Diaphanousness, Representationalism. In T. Alter & S. Walter (eds.), *Phenomenal Concepts and Phenomenal Knowledge: New Essays on Consciousness and Physicalism.* P. 52–64.

Kim, Jaegwon. (2001). Lonely Souls: Causality and Substance Dualism. In K Corcoran (Ed), *Soul, Body, and Survival: Essays on the Metaphysics of Human Persons.* P. 30-43. Cornell University Press: Ithaca, USA.

Leal, Manuel., & Powell, Brian J. (2011). Behavioral Flexibility and Problem Solving in a Tropical Lizard. In *Biology Letters*. Vol. 8. P. 28-30. The Royal Society. London, UK.

Leibniz, Gottfried. (1714). *The Monadology*. Trans. By Robert Latta. Oxford University Press. Oxford, UK.

McGovern, K., & Baars, B.J. (2007). Cognitive theories of consciousness. In Philip Zelazo, Morris Moscovitch, & Evan Thompson (Eds.), *The Cambridge Handbook of Consciousness.* P. 177-206. Cambridge University Press. Cambridge, UK.

Morin, Roc. (2015). *A Conversation with Koko the Gorilla*. The Atlantic. Washington, USA.

Nagel, T.. (1974). What Is It Like to Be a Bat?, in *The Philosophical Review.* Vol 83(4). P. 435-450. Duke University Press. Durham, USA.

Pennartz et al. (2019). Indicators and Criteria of Consciousness in Animals and Intelligent Machines: An Inside-Out Approach, in *Frontiers in Systems Neuroscience.* Vol 13:25. P. 1-23. Frontiers Media. Lausanne, CH.

Pierson, Lee M., & Trout, Monroe. (2005). What is Consciousness For?, in *New Ideas in Psychology.* Vol. 47. P. 62-71. Pergamon Press. Oxford, UK.

Price, Mark C., & Norman, Elisabeth. (2008). Intuitive Decisions on the Fringes of Consciousness: Are They Conscious and Does it Matter?, in *Judgment and Decision Making.* Vol. 3:1. P. 28-41. Society for Judgment and Decision Making.

Rosch, E. H. (1973). Natural Categories, in *Cognitive Psychology.* Vol. 4. P. 328-350. Elsevier. Lausanne, CH.

Rosenthal, David M. (1986). Two Concepts of Consciousness, in *Philosophical Studies.* Vol. 49. P. 329-359. D. Reidel Publishing Company. Dordrecht, NL.

Searle, John. (1980). Minds, Brains, and Programs. In *Behavioral and Brain Sciences.* Vol. 3:3. P. 417-457. Cambridge University Press. Cambridge, UK.

Sellars, Wilfrid. (1962). Philosophy and the Scientific Image of Man. In R. Colodny (Ed.) *Frontiers of Science and Philosophy.* P. 35-78. University of Pittsburgh Press. Pittsburgh, USA.

Stoerig, Petra. (2007). Hunting the Ghost. In Philip Zelazo, Morris Moscovitch, & Evan Thompson (Eds.), *The Cambridge Handbook of Consciousness*. pp. 707-730.Cambridge University Press. New York, USA.

Strawson, P. F. (1959). *Individuals.* Routledge Press. London, UK.

Taylor et al. (2011). New Caledonian Crows Learn the Functional Properties of Novel Tool Types. In *Plos One.* Vol. 6:12. Cambridge, UK.

Tononi, Giulio., & Koch, Christof. (2015). Consciousness: Here, There, and Everywhere? In *Philosophical Transactions of the Royal Society B.* Vol. 370:1668. London, UK.

Thagard, Paul. (1996). *Mind: Introduction to Cognitive Science.* The MIT Press. Cambridge, MA, USA. P.124.

Tye, Michael. (1995). *Ten Problems of Consciousness: A Representational Theory of the Phenomenal Mind.* The MIT Press. Cambridge, USA.

Tye, Michael. (2003). *Consciousness and Persons.* MIT Press. Cambridge, USA.

Tye, Michael. (2019). *Tense Bees and Shell Shocked Crabs: Are Animals Conscious?.* Oxford University Press. Oxford, UK.

Wittgenstein, Ludwig. (1953). *Philosophical Investigations*. Trans. By G. E. M. Anscombe. Basil Blackwell Ltd. Oxford, UK.

Zeki, S. (2003). The disunity of consciousness. In *Trends in Cognitive Science.* Vol.7:5. P. 214-218. Elsevier. USA.

Zlatev, J., Racine, T. P., Sinha, C., & Itkonen, E. (Eds.). (2008). *Converging evidence in language and communication research (CELCR): Vol. 12. The shared mind: Perspectives on intersubjectivity*. John Benjamins Publishing Company. USA.