MONASH University

The Problem of AI Consciousness

Goancheol Shin

Bachelor of Arts with Honours

A thesis submitted for the degree of Master of Arts at

Monash University in 2023

School of Philosophical, Historical, and International Studies

Copyright notice

I certify that I have made all reasonable efforts to secure copyright permissions for third-party content included in this thesis and have not knowingly added copyright content to my work without the owner's permission.

Abstract

This thesis examines what I call the *problem of AI consciousness*: how do we tell that an AI is conscious? At its core, this question is just an extension of a long-standing philosophical problem called the *problem of other minds* (POM): how do I know that other people have minds? Although the POM is in many ways out of fashion, my view is that it provides us with useful tools for both thinking about and answering various questions we have about the distribution of consciousness. I will take three solutions that have been proposed to the POM – the argument from analogy (AA), the argument from best explanation (ABE), and the argument from criteria (AfC) – and apply them to the problem of AI consciousness.

The AA holds that we are justified in believing that an AI system is conscious if it exhibits the kinds of behaviour that are reliably correlated with conscious experiences in each of our own cases (or simply, in human cases). This argument is widely regarded as an inadequate solution that is superseded by the ABE. I suggest, however, that the common objections raised against it can be successfully countered. The current orthodoxy is the ABE which draws its strengths from sharing the same abductive framework that has proven crucial to the modern scientific method. This argument suggests that we are justified in believing that an AI system is conscious if its behaviour is best explained by the hypothesis that it is conscious.

Although there are numerous attractive elements to the AA and the ABE, I argue that the core challenge facing them is their underlying assumption regarding our concept of consciousness – namely, that we can think about consciousness without any reference to behaviour. I argue, on Wittgensteinian grounds, that there is a crucial *conceptual* link between consciousness and behaviour, and that any plausible solution to the problem of AI consciousness must not only recognize but cohere with this idea. The AfC, in my view, succeeds in this manner and therefore is my favoured solution. According to this argument, certain behavioural patterns are our criteria for the concept CONSCIOUSNESS – that is, they are the very means by which grasp the concept. On this view, we are justified in believing that an AI system is conscious if its behaviour satisfies such criteria.

Declaration

This thesis contains no material which has been accepted for the award of any other degree or diploma at any university or equivalent institution and that, to the best of my knowledge and belief, this thesis contains no material previously published or written by another person, except where due reference is made in the text of the thesis.

Signature: ……

Print Name: Goancheol Shin

Date: 07/06/2023

Acknowledgements

I am grateful for all the help and support I received in writing this work. I would like to give my sincerest thanks first to Tim Bayne for sharing his knowledge and providing me with kind-hearted support. A very special thank you also to Monima Chadha who always went out of her way to help me – your constant encouragement has meant a lot to me.

I cannot list them all, but I thank each and every one of my friends, both those here at Monash with me and those scattered elsewhere. Special thanks to those who have given me valuable feedback on my writing. I also thank the authors to whom I reached out for clarification on their work – their replies were not only helpful but a warm surprise. Last but not least, thank you to my family!

This research was supported by an Australian Government Research Training Program (RTP) Scholarship.

Table of Contents

Chapter 1: Introduction

From disembodied Artificial Intelligence (AI) systems like J.A.R.V.I.S from *Iron Man* and Samantha from *Her*, to the eponymous trash-compacting robot from *WALL-E* and the humanoid Ava from *Ex Machina*, when we watch a science-fiction film featuring intelligent machines – especially the robotic or, better yet, humanoid kind – we find ourselves naturally imagining in them an inner world of thoughts, emotions and sensations much like our own. Of course, few, if any, believe that these fictional characters are conscious – they are surely nothing more than computer-generated imagery represented by pixels on a screen. But fictional depictions of sophisticated machines provide an imaginative entry into fascinating questions about the distribution of mental phenomena in our world. Can machines think? Can they have emotions? Can they feel pain? In a world of accelerating technological advancement, we are left wondering whether these scenarios may soon become reality.

In fact, in recent decades, there has been growing interest in the development of conscious machines. This interest is no longer one of mere speculation but of systematic pursuit involving theoretical research as well as modelling and robotics projects.[1] This emerging field of artificial consciousness (or machine consciousness) has two primary research agendas. 'Strong' artificial consciousness research aims to create artificial systems that are conscious in

[1] See, for example, James A. Reggia, Garrett E. Katz, and Gregory P. Davis, 'Artificial Conscious Intelligence', *Journal of Artificial Intelligence and Consciousness* 7, no. 1 (2020): 95–107; James A. Reggia, 'The Rise of Machine Consciousness: Studying Consciousness with Computational Models', *Neural Networks* 44 (2013): 112–31; David Gamez, *Human and Machine Consciousness* (Open Book Publishers, 2018).

the sense in which you and I are.[2] It should be no surprise that there are no credible success stories to date here.[3] On the other hand, the goal of 'weak' artificial consciousness research is not to create conscious systems, but to better understand human or biological consciousness through methods found in computer science and AI – e.g., computational models of mechanisms associated with consciousness. According to this approach, we are no more likely to find genuine feelings and emotions within a 'consciousness-simulation' than find actual radiation emitting from simulated gamma-rays or wetness in fluid dynamics simulations.[4] Although I do not engage with the details of either form of research in this thesis, my aims are tied more closely to strong artificial consciousness research.

Let me now introduce an imaginary AI system to help give body to our discussion ahead. The system I am imagining is a humanoid robot that strikingly resembles the way we humans look and behave across a wide range of contexts. He is also equipped with an advanced language model and a human-like voice synthesizer.[5] To be sure, he is not made of exactly the same materials as us or other biological creatures – he is a silicon-based, rather than carbon-based, system – and as a result does have rather unique preferences when it comes to food, choosing films, and the like. But he is a remarkable creation nonetheless. I will name him, *Replicant*, in honour of the fictional humanoids from the 1982 film, *Bladerunner*. I make frequent reference to Replicant throughout my thesis, so keep him in mind.

Suppose further that Replicant is not remotely controlled and, to defer certain objections for the moment, that his internal mechanisms are as complex as ours, interacting with inputs as well as with each other to generate his behaviour. In the world of fiction, our projection of conscious life to its characters, although useful, is most certainly nothing more than fanciful imagination on our part. But what would we say about Replicant – a real-life humanoid that looks and behaves like us in real-time? How, if at all, does your intuition differ?

My hunch is that you will not be so sceptical anymore. I suspect much of the reason we tend to think that present-day AI systems – be they a computer, an operating system, or a robot – are not conscious is that they are not complex enough, both in terms of their internal mechanism and behaviour. As Michael Tye points out, we tend to think that "robots by their

[2] Anil Seth, 'The Strength of Weak Artificial Consciousness', *International Journal of Machine Consciousness* 1, no. 1 (2009): 71–82.
[3] That said, some are worried about the possible outcomes of this research. Thomas Metzinger, for one, has called for global moratorium on artificial consciousness research. Thomas Metzinger, 'Artificial Suffering: An Argument for a Global Moratorium on Synthetic Phenomenology', *Journal of Artificial Intelligence and Consciousness* 8, no. 1 (2021): 43–66.
[4] Of course, whether this is right – whether a fine-grained digital simulation of the brain's functional organization cannot produce consciousness – is a separate question.
[5] For convenience, I use the he/him/his pronoun for Replicant (rather than 'it').

nature are unemotional beings who do just what they are programmed to do," and we tend to have an "oversimplified picture of how [they] would behave."[6] But these reasons carry little weight in Replicant's case since he is far more complex than any of today's robots.

Now, whatever our intuitions are now or in the future, the important question is whether we are justified in the view we take. My intention for introducing the above train of thoughts is first to lead us to the following more general question: on what basis can I justifiably believe that any creature other than myself is conscious? Safe to say, you believe that other humans are conscious. In fact, the reasons we have for feeling as if fictional AI systems are conscious seems likely to be tied to this fundamental belief. But what is it about our fellow humans that justifies this belief? How can I be sure that other humans are not merely 'philosophical zombies' – creatures that are physically and functionally similar to me but lack consciousness? This question of how we can justifiably believe (or know) that other human beings have minds has been called the *problem of other minds* (POM).[7] I will introduce this problem and those closely related to it in more detail soon.

Second, having considered these questions, I want us to trace our steps back and use our answers as tools for addressing the same sort of curiosity about AI systems: how might we be justified in believing that an AI is (or is not) conscious? I will call this the *problem of AI consciousness*.[8] More precisely, I am interested in how three often discussed solutions to the POM – the argument from analogy (AA), the argument from best explanation (ABE), and the argument from criteria (AfC) – might help us with the problem of AI consciousness. I should be clear that the question I am interested in is not whether AI consciousness is *possible*, but rather how we might *tell* that an AI is conscious. The two questions are tied but should not be confused. Even if AI consciousness is possible, without the tools with which we can distinguish conscious systems from mere 'zombie' systems, we cannot determine whether we have created a conscious AI (intentionally or not).

The transition from human cases to AI systems is not straightforward. Even if we find a plausible answer for the human POM, we should not expect it to be straightforwardly applicable to AI systems. For example, we cannot rely on evolutionary proximity or neurobiological similarity in AI cases. Indeed, as Tim Bayne notes, "it is questions about

[6] Michael Tye, *Tense Bees and Shell-Shocked Crabs: Are Animals Conscious?* (Oxford University Press USA, 2016), 191.
[7] I prefer the term 'justified belief' over 'knowledge' because the latter intuitively implies certainty (which, in my view, is unnecessarily strong in the context of the problem of AI consciousness). Most writers in the POM literature, however, often use them interchangeably. I myself do the same at times.
[8] In some respects, 'artificial consciousness' is preferrable to 'AI consciousness' because AI research is not always relevant to the question of artificial consciousness (e.g., not all artificial systems need be an AI).

consciousness in AI systems that threaten to pose the most serious version of the problem of other minds."[9]

1.1 Focusing on Consciousness

Although I draw heavily on the tools within the POM debate, my interest in this thesis, as you may have noticed, lies with *consciousness* rather than the *mind*. The notion of mind is an important – perhaps even indispensable – part of both philosophy of mind and the 'mind sciences', and it is just as important in our everyday vocabulary. Why do I, then, focus on consciousness?

One reason has to do with the vagueness of the notion of mind. We all have an intuitive grasp of what a mind is, but when you carefully ponder upon what it is, what do you find? Rather than finding a clear, unified target, I suspect you will end up with a rather fuzzy collection of various kinds of mental phenomena: sensations, emotions, thoughts, memories, beliefs, and more. Of course, there are difficulties in pinning down each of these phenomena too, but it certainly does not make it easier to think of them all at once with a broad brush. Each mental phenomenon presents unique questions and challenges that resist uniform treatment often given to them by the POM literature.

This is especially true when considering AI systems. We tend to assume that other human beings (and even many non-human animals) have minds with little worry around what we mean by that term, but the same cannot be said about AI systems, for we are unsure whether they are minded beings and, even if they are, whether the mental states they harbour are anything like our own. Perhaps a particular AI system can be said to possess beliefs and desires but not sensations or emotions, whereas another may experience feelings, but have no thoughts. Is one a mind but not the other? What is the most important ingredient in a mind? In this way, asking whether an AI has a mind makes it all too easy to overlook these important differences, as well as lure us into further difficult questions. Hence, although some mental phenomena may be intimately related, it will be helpful to focus on a particular phenomenon. This will allow me to arrive at more cohesive answers to the questions I will be exploring (or at least make the path more manageable). I will still be using the terms 'mind' and 'POM', but here onwards, I really just mean 'consciousness' by them (unless, of course, specified otherwise).

[9] Tim Bayne, *Philosophy of Mind: An Introduction* (New York: Routledge, 2021), 212.

A more positive reason for my focus on consciousness is that it is a phenomenon that is intimately familiar to each of us. Our waking life is filled with rich waves of conscious experiences. In fact, I suspect that when each of us dwells on the nature and distribution of the mind, we naturally gravitate towards thinking about conscious agents and about conscious experiences. Moreover, consciousness is regarded by many philosophers and scientists as one of the most important aspects of the mind. Indeed, it is at the heart of many lively debates in the philosophy of mind with far-reaching implications. Exploring the problem of AI consciousness is therefore not only fascinating in its own right, but also fits in well with current philosophical trends. On a related note, we have also seen that artificial consciousness is an emerging research field. Hence, the contents of this thesis can be seen as having a direct target. Overall, the problem of AI consciousness is an increasingly important conversation to have.

At the same time, consciousness is a most puzzling thing. It has been characterized and studied in a countless number of ways and there are many ongoing disputes regarding the plausibility of each view. Distinctions of the types of consciousness have been made across lines of phenomenality, access, reflexivity, transitivity, representation, and more. [10] Distinctions also exist along what we might call structural lines, between, for example, 'levels' of consciousness, global and local states of consciousness, and specific and generic (or state and creature) consciousness.[11] Then there are disputes about the features of consciousness including those about subjectivity, phenomenality, intrinsicality, unity, and intentionality.[12] Matters are further complicated by the fact that philosophers and scientists often use different labels to mean more or less the same thing as 'consciousness' – e.g., 'awareness', 'experience', 'qualia', and 'sentience'. To boot, the same label is often used to mean different things.

However, perhaps the most serious overarching difficulty for the study of consciousness has to do with the metaphysical nature of consciousness. Is consciousness a physical phenomenon? What is its causal status? What is its relation to the brain, cognition, and behaviour? These questions are difficult to answer because consciousness at its core does

[10] For the phenomenal/access distinction, see Ned Block, 'On a Confusion about a Function of Consciousness', *Behavioral and Brain Sciences* 18, no. 2 (June 1995): 227–47. For an overview of representationalism, see William Lycan, 'Representational Theories of Consciousness', in *The Stanford Encyclopedia of Philosophy*, ed. Edward N. Zalta, Fall 2019 (Metaphysics Research Lab, Stanford University, 2019), https://plato.stanford.edu/archives/fall2019/entries/consciousness-representational/.

[11] See, for example, Tim Bayne, Jakob Hohwy, and Adrian M. Owen, 'Are There Levels of Consciousness?', *Trends in Cognitive Sciences* 20, no. 6 (2016): 405–13; Andy Kenneth Mckilliam, 'What Is a Global State of Consciousness?', *Philosophy and the Mind Sciences* 1, no. 2 (2020).

[12] For an overview of these distinctions, see Section 4 in Robert Van Gulick, 'Consciousness', in *The Stanford Encyclopedia of Philosophy*, ed. Edward N. Zalta, Winter 2021 (Metaphysics Research Lab, Stanford University, 2021), https://plato.stanford.edu/archives/win2021/entries/consciousness/..

not seem amenable to physical (i.e., structural and functional) explanation in the way many other phenomena are (e.g., combustion, disease, gravity, and life). In turn, these metaphysical puzzles raise serious methodological hurdles for scientifically studying consciousness. How do we, for example, distinguish between a neural state or process that is merely correlated to consciousness and that which is causally related to it?[13] Many ethical questions turn on these puzzles too. For instance, how do we ensure that a patient under general anaesthesia is unconscious rather than merely led to forget a fully conscious surgery (i.e., distinguish between an anaesthetic and a paralytic amnestic)?[14]

These metaphysical considerations also have direct relevance for both the possibility of AI consciousness and the problem of AI consciousness, some of which I will touch on in the coming chapters. For example, if consciousness is intimately tied to biological properties or processes, we have reason to doubt that AI consciousness is possible at all, for artificial systems are likely to be made of synthetic materials. By contrast, if consciousness is a functional property or arises by virtue of organizational (or computational) complexity, then we have reason to be more optimistic since the same level of complexity may be achieved on synthetic substrates. As out of fashion as it may be, if (logical) behaviourism was true – i.e., being conscious is just a matter of behaving or being disposed to behave in a certain way – then we should have little reason to doubt the possibility of conscious AI (and the problem of AI consciousness would hardly even arise).[15] In sum, reflecting on these metaphysical questions are relevant for both raising and answering the POM and the problem of AI consciousness since it gives us an idea of *what* it is that we are wondering the target subjects have.

That said, my focus in this thesis is not on the metaphysics of consciousness and I will be side-stepping as much of the aforementioned debates as possible, bringing them up only where directly relevant to my concerns.[16] It should be helpful to note here that none of the AA, the ABE, and the AfC, at least as I present them, has any rigid metaphysical commitment, and as such any one of them may be endorsed by competing metaphysical views. More broadly, it is worth adding that one's view on the nature of consciousness need not limit the kinds of data

[13] Jakob Hohwy and Tim Bayne, 'Causes, Confounds and Constituents: The Neural Correlates of Consciousness', in *The Constitution of Phenomenal Consciousness: Toward a Science and Theory,* ed. Steven M. Miller, Advances in Consciousness Research (John Benjamins Publishing Company, 2015), 155–76.
[14] Michael T. Alkire, Anthony G. Hudetz, and Giulio Tononi, 'Consciousness and Anesthesia', *Science (New York, N.Y.)* 322, no. 5903 (2008): 876–80.
[15] It is worth noting, however, that this does not necessarily mean that the nature of the material basis or internal mechanism is irrelevant, since they may have an impact on the kind of behaviour that a system can generate.
[16] E.g., in Section 2.2.3 where I evaluate challenges to the value of behavioural evidence, in Sections 3.2 and 3.3 where the causal profile of consciousness enters into doubts about the ABE, and in several places in Chapter 4 where the phenomenalist intuition is shown to be relevant to the way we can think and talk about consciousness.

one can accept as evidence of consciousness. For instance, you certainly do not have to be a behaviourist to think that behaviour is useful evidence of another's consciousness. Even an epiphenomenalist who believes that consciousness is causally inert could make use of behaviour as evidence – say, if it accurately tracks conscious states.

What do I, then, mean by 'consciousness'? Here, I want to avoid getting caught up in metaphysical disputes by focusing on the pre-theoretical grasp we have of consciousness.[17] Of course, our pre-theoretical intuitions are fallible, and I am wary of the fact that they are about *human* consciousness. But I will simply have to take this as my starting point. Thus, for my purposes, I characterize consciousness as the state which we enjoy while we are awake and is typically contrasted with the absence of consciousness in dreamless sleep or the loss of consciousness following a traumatic head injury or death.[18] When we are awake and going about our day, we possess a subjective perspective from which our experiences sprout. There is *something it is like* to be us – the 'lights are on', so to speak.[19] Perhaps this light is not switched on all the time during wakefulness or not always on at the same level of intensity. For example, it may be brighter when we stub our toes or reflect on consciousness itself than when we are washing the dishes or drifting off to sleep. We can also speak of different 'colours' of conscious experience: the taste of lemon, a light static shock, and the sudden sense of losing balance are surely distinct conscious experiences despite sharing a similarly 'sharp' or sudden character. Nonetheless, as a whole, I think we can agree that we enjoy a special kind of state unavailable to us while we are unconscious.

1.2 The Epistemic Problem of Other Minds

Whenever the phrase 'the problem of other minds' is mentioned, a common assumption is that there is a single, unified issue being discussed:

The Epistemic Problem: How do I justify the belief that others are conscious?

This is arguably the most natural philosophical question we have about our knowledge (or belief) of other minds, and it should not be a surprise that it has a rich philosophical history.

[17] Another useful step I will be taking throughout the thesis is to focus on a particular conscious state or a set of conscious states as a proxy for consciousness. My primary example is pain.
[18] There is some question as to whether there may be some level of conscious awareness during dreamless sleep. I will bracket such concerns. Jennifer M. Windt, Tore Nielsen, and Evan Thompson, 'Does Consciousness Disappear in Dreamless Sleep?', *Trends in Cognitive Sciences* 20, no. 12 (2016): 871–82.
[19] Thomas Nagel, 'What Is It Like to Be a Bat?', *The Philosophical Review* 83, no. 4 (1974): 435–50.

René Descartes is perhaps the most notable early figure in the Western tradition to raise this question:

> But then if I look out of a window and see men crossing the square, as I just happened to have done, I normally say that I see the men themselves. . . Yet do I see more than hats and coats which could conceal automatons?[20]

The epistemic problem therefore has the right of being called the *traditional* POM. This is the problem I have been referring to by the term 'POM', and unless stated otherwise, this will remain the same.

Now, although the POM has been discussed primarily within philosophical contexts and with respect to other (neurotypical) human beings, we often engage with what are very similar questions in everyday and scientific scenarios. A child might wonder whether the bug they are poking *feels* it, whether the dolphins in the aquarium are *happy*, or whether their teddy bear is '*alive*'. Neuroscientists study whether persistent vegetative state patients are aware (rather than merely wakeful).[21] Those interested in the abortion debate take the question of prenatal sentience very seriously.[22] Animal scientists and ethicists are curious about which non-human animals, especially among those evolutionarily distant from us (e.g., invertebrates like insects and octopuses), are sentient.[23]

The POM can be seen as just an extension (and formalization) of these types of curiosity we have about the creatures that we share our world with. The common goal is to strike a balance between being overly conversative in our estimations so as to fail to attribute consciousness where it is due (false negatives) and being too liberal so as to misattribute consciousness (false positives). That said, there is a difference worth noting here. The POM

---

[20] John Cottingham, Robert Stoothoff, and Dugald Murdoch, *The Philosophical Writings of Descartes: Volume 2* (Cambridge University Press, 1984), 21. That said, Descartes is certainly not the first to notice this problem. Saint Augustine (354–430), for one, was deeply interested in the problem. For the Western history of the POM, see preface, introduction, and Chapters 1 and 2 in Anita Avramides, *Other Minds* (Routledge, 2000). In the Eastern tradition, 7th century AD Indian Buddhist philosopher Dharmakīrti is said to be "perhaps the first ever thinker to make a systematic attempt to come to grips with this problem." Ramesh Kumar Sharma, 'Dharmakīrti on the Existence of Other Minds', *Journal of Indian Philosophy* 13, no. 1 (1 March 1985): 55.

[21] Adrian M. Owen and Martin R. Coleman, 'Detecting Awareness in the Vegetative State', *Annals of the New York Academy of Sciences* 1129 (2008): 130–38.

[22] Hugo Lagercrantz and Jean-Pierre Changeux, 'The Emergence of Human Consciousness: From Fetal to Neonatal Life', *Pediatric Research* 65, no. 3 (March 2009): 255–60.

[23] Peter Godfrey-Smith, *Other Minds: The Octopus, the Sea, and the Deep Origins of Consciousness*, First edition (New York: Farrar, Straus and Giroux, 2016); Colin Allen and Michael Trestman, 'Animal Consciousness', in *The Stanford Encyclopedia of Philosophy*, ed. Edward N. Zalta, Winter 2020 (Metaphysics Research Lab, Stanford University, 2020), https://plato.stanford.edu/archives/win2020/entries/consciousness-animal/; Morten Overgaard, 'Insect Consciousness', *Frontiers in Behavioral Neuroscience* 15 (2021).

has been traditionally motivated by the threat of solipsism, the view that I am the only conscious being in the universe. The solipsist challenges not merely our actual belief in another's consciousness, but our very capacity to ever justifiably believe in it. I will call this the *philosophical* POM.

By contrast, our everyday and scientific queries about the identification and overall distribution of consciousness do not share this solipsistic concern. Instead, the assumption here is that we can wonder or even have doubts about the presence of consciousness in a creature without holding that there is a problem regarding our belief in other conscious creatures more generally. After all, why should the possibility that insects are not conscious cast doubts about dolphin consciousness? In fact, attempts to identify consciousness in non-human animals or contested (or 'marginal') human cases typically proceed on the assumption that neurotypical humans are conscious. I will call this the *scientific* POM.

The problem of AI consciousness is most naturally viewed as a scientific POM. Indeed, researchers that are attempting to create conscious systems or develop tests of artificial consciousness often model their approach on the human brain based on the assumption that human beings are conscious. Now, although the three frameworks I will be exploring – the AA, the ABE, and the AfC – have their root in the philosophical POM, I see no reason to think that they cannot help us with the scientific POM. The two problems stem from different assumptions and goals, but the overall nature and structure of their questions are the same. To my eyes, the two sets of problems are continuous.

Before returning to the explanation of the epistemic POM, I want to draw one more distinction – between the *specific* and the *generic* POM.[24] The specific problem concerns our knowledge of the *content* of another's consciousness – i.e., *what* they are experiencing. Could it be that others perceive colours in very different ways than I do? How do I ensure that others are not frequently lying to me about their thoughts and feelings? The generic problem concerns how we know that another is a conscious being *at all* (rather than a philosophical zombie). When it comes to AI systems, we are dealing primarily with the generic problem since we know very little about the possibility of AI consciousness.

[24] My terminology here borrows from the consciousness literature which distinguishes between specific consciousness (toothache, smell of coffee) and generic consciousness (the property of being conscious rather than non-conscious). Anita Avramides puts the same distinction in terms of the *thin* versus the *thick* problem. A. Avramides, 'On Seeing That Others Have Thoughts and Feelings', *Journal of Consciousness Studies* 22, no. 1–2 (2015): 138–55. Mark Addis puts it in terms of the moderate versus the radical problem. Mark R. Addis, *Wittgenstein: Making Sense of Other Minds* (Ashgate, 1999), 2.

Now, why precisely does the epistemic problem – both in the philosophical and scientific contexts – arise? I do not wonder how I might know that your eyes are brown, nor would I find difficulty in figuring out your height. Why do I wonder, then, whether you are conscious? Åsa Wikforss captures the answer well:

> It is widely held that there is a deep-lying asymmetry between first- and third-person knowledge: whereas we know our own minds *directly*, knowledge of other minds is always *indirect* or dependent on epistemic intermediaries. The sceptical worries about other minds feed on this idea since in comparison with the directness of self-knowledge, knowledge of other minds seems precarious.[25]

Alec Hyslop gives us helpful examples to further illustrate the idea:

> We often know directly that we are in a certain mental state. Typical cases would be where we are in serious pain, are itching, are smelling a rose, seeing a sunflower, are depressed, believe that today is Tuesday, and so forth. We do not always know directly that we are in the mental state we are in but what is striking is that we never have direct knowledge that other human beings are in whatever mental state they are in. It is this stark asymmetry that generates the epistemological problem of other minds.[26]

I will call the claim that there is a deep asymmetry between first- and third-person knowledge of consciousness, the *asymmetry thesis*. Suppose, for example, I see my friend Jones step on a piece of Lego, and he reacts by jumping and yelling "ouch!" It is natural to believe that Jones is in pain and that, were I to make the same mistake, my experience would be similar. Yet, it is unclear how I am to validate this belief, for it seems I cannot directly know his conscious experiences. By contrast, there is no 'problem of other heights' because Jones' height is openly available to my senses – I can straightforwardly measure his height in more or less the exact way I can my own. The same goes for arguably any other of his physical features.

We should be absolutely clear about how deep the epistemic asymmetry runs. Suppose it is possible to feel pain located in someone else's body. After all, it is not inconceivable, at

[25] Åsa Wikforss, 'Knowledge, Belief, and the Asymmetry Thesis', in *Knowing Other Minds* (Oxford: Oxford University Press, 2019), 41.
[26] Alec Hyslop, 'Other Minds', in *The Stanford Encyclopedia of Philosophy*, ed. Edward N. Zalta, (Metaphysics Research Lab, Stanford University, 2014), https://plato.stanford.edu/archives/spr2019/entries/other-minds/.

least Ludwig Wittgenstein thought, that on the evidence of pain alone, that I radically mislocate my experience:

> Suppose I feel a pain which on the evidence of the pain alone, e.g., with closed eyes, I should call a pain in my left hand. Someone asks me to touch the painful spot with my right hand. I do so and looking round perceive that I am touching my neighbour's hand.[27]

Today, we not only know that this sort of error is possible, but we also know that it is not particularly difficult to induce them even in subjects with healthy brain function. In what is called the *body transfer illusion*, manipulating the visual perspective of human subjects can produce the illusion of conscious experiences in a body part that is not theirs (or in an inanimate object). Various forms of the rubber hand experiment, mirror therapy, and virtual reality experiences are all well-documented instances of this phenomenon.[28]

Although these cases are illusions after all (i.e., they are not genuine instances of experiencing something in another's body), they nonetheless give us some prima facie reason to think that one's own pain could be felt in another's body. To that extent, one might be led to think that we could, at least in principle, know (or be justified in believing) that Jones experiences pain. Yet this would be a mistake, for what we need to solve the POM (at least according to the sceptic) is not my having conscious experiences in Jones' body, but my knowing that these experiences are felt by Jones.

What about cases where pain appears to be shared with another person? One possible instance may be in conjoined twins.[29] Here, pain is presumed to be not merely qualitatively identical to that of another, nor just felt as located in their body, but spatio-temporally identical to theirs. Some have argued that this is a case of directly experiencing and knowing another's

[27] Ludwig Wittgenstein, *The Blue and Brown Books: Preliminary Studies for the 'Philosophical Investigations'* (Oxford, England: Harper & Row, 1958), 49.
[28] Matthew Botvinick and Jonathan Cohen, 'Rubber Hands "Feel" Touch That Eyes See', *Nature* 391, no. 6669 (1998): 756; Brenda L. Chan et al., 'Mirror Therapy for Phantom Limb Pain', *The New England Journal of Medicine* 357, no. 21 (2007): 2206–7; Mel Slater et al., 'First Person Experience of Body Transfer in Virtual Reality', *PLOS ONE* 5, no. 5 (2010): e10564.
[29] Alec Hyslop, *Other Minds*, Synthese Library, v. 246 (Dordrecht ; Boston: Kluwer Academic Publishers, 1995), 5–6; Tom Cochrane, 'A Case of Shared Consciousness', *Synthese* 199, no. 1 (2021): 1019–37. Or suppose 'mind-melding' through surgical thalamic bridging is made possible in the distant future.

conscious experiences.[30] However, the POM would remain (the sceptic claims), for one would still lack the guarantee from their perspective that their counterpart feels anything at all.

In sum, the crux of the problem is not that I cannot directly experience another's consciousness (although that is very contentious), but that I cannot directly know it *as such*, as belonging to *them*.[31] Alec Hyslop and Frank Jackson summarize the depth of the issue well:

> It is commonly asserted that what generates the problem of other minds is the impossibility of being directly aware of the mental states of others. This is a mistake. First, it is not at all clear that it is impossible to be directly aware another's mental state. Secondly, and more importantly, being directly aware of someone else's pain would not, in itself, tell me that that person was in pain, because I should still require grounds for supposing that what I was aware of was someone else's pain (as well as my own).[32]

At its core, the problem of AI consciousness – or, for that matter, any question regarding our knowledge of another's consciousness – is generated by this apparent epistemic asymmetry.

1.3 The Conceptual Problem of Other Minds

The epistemic problem dominated the historical debate around other minds. However, around the middle of the 20th century, philosophers began to recognize a very different way of questioning our knowledge of other minds:

> *The Conceptual Problem*: How can I so much as *conceive* of consciousness in others?

If Descartes was the (Western) champion of the epistemic problem, Wittgenstein is perhaps the first to bring the conceptual problem to light:

[30] William Hirstein, 'Mindmelding: Connected Brains and the Problem of Consciousness', *Mens Sana Monographs* 6, no. 1 (2008): 110–30; William Hirstein, 'Sharing Conscious States', in *Mindmelding: Consciousness, Neuroscience, and the Mind's Privacy*, ed. William Hirstein (Oxford University Press, 2012), 148–64.
[31] Hyslop, *Other Minds*, 1995, 55.
[32] A. Hyslop and F. C. Jackson, 'The Analogical Inference to Other Minds', *American Philosophical Quarterly* 9, no. 2 (1972): 168.

> If one has to imagine someone else's pain on the model of one's own, this is none too easy a thing to do: for I have to imagine pain which I *don't feel* on the model of the pain which I *do feel*.[33]

I will explore the conceptual problem in greater depth in Chapter 4 where it is most relevant, but it will be helpful to take the time now to sketch the core issue.

Here, Wittgenstein questions how the possibility of pain in others can be entertained at all in the face of the natural assumption that the very concept of pain that I have stems from my own pain-experiences. The issue here is not whether I can be justified in believing that others experience (or do not experience) pain – that would be an *epistemic* problem. Rather, the issue is how it is that third-person attributions of pain are *intelligible* (or meaningful) at all. I will characterize the problem as that of a tension that arises between two intuitively plausible claims: (1) grasping certain mental concepts such as CONSCIOUSNESS and PAIN requires introspective attention to one's own mental states, and (2) it is possible to intelligibly apply such concepts to others.

Claim 1 seems secure – it is both widely endorsed (at least implicitly) and intuitively attractive. Claim 2 also seems secure: we do, as a matter of plain fact, attribute mental states to other human beings and various non-human animals (and deny them to rocks and water bottles). But there is a tension here between the apparent privacy of our mental concepts and the natural generality with which we extend them to others. How is it possible to apply what appears to be an inherently private concept – a concept that is based on and refers to something which only I have direct knowledge of – to others whose mental states I have no comparable access? All I can see is surely just their behaviour (or, at best, their brain states).

The main relevance of the conceptual problem for the problem of AI consciousness derives from the fact that wondering whether an AI is conscious requires that the mental concepts we have at our disposable are intelligibly applicable to others. Hence, the conceptual problem, as Anil Gomes writes,

> appears more basic than any concern about knowledge of other minds, for knowledge claims presuppose conceptual capabilities, and thus any problem with accounting for

[33] Ludwig Wittgenstein, *Philosophical Investigations*, ed. P. M. S. Hacker and Joachim Schulte, 4th Edition (Wiley-Blackwell, 2009), §302. There is, however, evidence of a discussion about the conceptual problem in Saint Augustine (354–430). See 'BRIEF INTERLUDE: ST AUGUSTINE' in Chapter 2 of Avramides, *Other Minds*, 2000.

> our ability to think about other minds will, a fortiori, pose a problem for explaining the possibility of our knowing about them.[34]

Both the epistemic and the conceptual problem were popular topics within philosophy throughout the twentieth century. Towards the end of the century, however, interest waned, and the POM as a whole became "one of the hallowed, if nowadays unfashionable, problems in philosophy."[35] As Jerry Fodor recalls in 1994, philosophy of mind during his time at graduate school (1956–60) "had two main divisions: the mind/body problem and the problem of other minds. . . Philosophical fashions change. It's gotten harder to believe that there is a *special* problem about the knowledge of other minds (as opposed to other anything elses)."[36] Fodor's claim is that the POM was gradually seen by philosophers at the time as just another species of scepticism that can be raised in any domain of knowledge, like that about the external world or the past.

Given that the problem of AI consciousness is not motivated by solipsistic or sceptical concerns (i.e., it is a *scientific* POM), the decline in the (philosophical) POM's popularity is a largely separate issue. That said, my view is that the POM does offer us useful tools for thinking about and tackling the problem of AI consciousness. The AA, the ABE, and the AfC are all examples of this. Further, many of the original objections to these arguments can and should be raised in the context of the scientific POM too.

1.4 The Psychological Problem of Other Minds

Some philosophers have suggested that there has been a revival of interest in our knowledge of other minds in recent years.[37] At the heart of this resurgence is a third way of examining our relation to other minds:

> *The Psychological Problem*: What kind of psychological process *underlies* our attribution of consciousness to others?

[34] Anil Gomes, 'Is There a Problem of Other Minds?', *Proceedings of the Aristotelian Society* 111 (2011): 356.
[35] Hyslop, 'Other Minds', 2014.
[36] Samuel Guttenplan, *A Companion to the Philosophy of Mind* (Cambridge: Blackwell, 1994), 292. The mind-body problem is the problem of how, if at all, mental states relate to physical states. For more on why the POM has declined in interest, see Gomes, 'Is There a Problem of Other Minds?'
[37] Anita Avramides, 'Other Minds', in *The Stanford Encyclopedia of Philosophy*, ed. Edward N. Zalta (Metaphysics Research Lab, Stanford University, 2020), https://plato.stanford.edu/archives/win2020/entries/other-minds/.

The question here is not whether I can be justified in believing that others are conscious (the epistemic problem), or how I can conceive of consciousness in others (the conceptual problem). Instead, the question is *how* (and *why*), in terms of psychological or cognitive mechanism, we attribute consciousness to others.[38] Asking how, for example, children learn to make psychological attributions to others, or why it is that we feel that fictional AI systems are conscious are questions of this type. We can liken these questions to those about what kind of developmental mechanisms underlie and affect personality traits, or what kind of physiological mechanisms are responsible for digestion or respiration.

The psychological problem is not a direct concern in my thesis, but it would be a mistake to dismiss it as entirely irrelevant to my concerns. For one, the boundary between the psychological and the epistemic problem can be blurry at times.[39] Take, for example, the *perceptual account of other minds*, according to which we can, at least sometimes for some mental states, perceive that others have them.[40] Although I do not discuss this account, it provides a good illustration of how the two problems can interact.[41] As Bayne writes, "The perceptual account is easily motivated, for we certainly talk as though we have perceptual access to mental phenomena."[42] We often say that we can see in someone's eyes that they are sad, that we like someone's way of thinking, or that we can see that one is in a lot of pain. Now, this account is often introduced in epistemic terms, as a justification for believing that others are conscious: I am justified in believing that Jones is in pain simply because I can see that he is in pain. But it is natural to view it simultaneously as a psychological account, as an explanation of which mechanism underlies our capacity to make third-person mental attributions (namely, a perceptual mechanism).[43] One might reason thus that our response to the problem of AI consciousness is likely to be (or perhaps must be), in one way or another, underscored by this perceptual mechanism.

---

[38] It is hence also called the *processing problem*. Matthew Parrott, 'Enquiries Concerning the Minds of Others', in *Knowing Other Minds*, ed. Anita Avramides and Matthew Parrott (Oxford University Press, 2019), 1–19.
[39] Indeed, the psychological POM literature often does not distinguish between the question of how we can legitimately know that others have mental states, and how or why, *psychologically*, we claim to know that others have mental states. Whether or not it is intentional, this ambiguity is unhelpful in my view.
[40] See Chapter 5: Other Minds in Quassim Cassam, 'The Possibility of Knowledge', in *The Possibility of Knowledge*, ed. Quassim Cassam (Oxford University Press, 2007); Avramides, 'On Seeing That Others Have Thoughts and Feelings'. Avramides. I should stress that the idea is not that we can see another's mental states, but see (the fact) *that* they have them.
[41] It is worth mentioning that the perceptual account is seen by some writers, rightly so in my view, as intimately tied to the AfC (see Section 4.3.2). Essay 14 in John McDowell, *Meaning, Knowledge, and Reality* (Cambridge, MA: Harvard University Press, 1998); P. M. S. Hacker, *Wittgenstein: On Human Nature* (Phoenix, 1997).
[42] Bayne, *Philosophy of Mind*, 197.
[43] Joel Krueger and Søren Overgaard, 'Seeing Subjectivity: Defending a Perceptual Account of Other Minds', in *Seeing Subjectivity: Defending a Perceptual Account of Other Minds* (De Gruyter, 2013), 297–320.

The ABE which will be explored in Chapter 3 is presented as an epistemic account, but it can also be seen in a psychological light. As Robert Pargetter introduces it:

> What is the nature of the inferences that we all so commonly, and rightly, make from certain behavioural evidence to the mental lives of other people? This paper explores features of the thesis that these inferences should best be viewed as being common scientific or hypothetic inferences, or arguments to the best explanation.[44]

The use of this form of inference (which I call *abductive* inference in Chapter 3) in a psychological account is better known as *theory-theory*. Its central claim, as the name suggests, is that the attribution of mental states to others involves the use of an implicit (folk-psychological) theory that best explains and predicts their behaviour.[45]

The AA, to be explored in Chapter 2, has a psychological counterpart as well: *simulation theory*. According to this view, we attribute mental states to others by imagining or 'simulating' their experiential perspective. The simulation theory dovetails nicely with the AA, for both take facts about one's own mental states as crucial to their respective proposal. The AfC too has been suggested as "perhaps best viewed as epistemological anthropology, describing how human beings, as a matter of psychological fact, treat other minds."[46] Now, how we, as a matter of psychological mechanism, attribute mental states to others may say little about the legitimacy of the attributions (that seems to be a fallacious appeal to nature), but it seems reasonable to think that a feature of a plausible psychological explanation is its fit with a plausible epistemic explanation (and vice versa).

There are also aspects in which the psychological problem is intimately tied to the conceptual problem. Again, if the mechanism by which we learn to attribute mental states to others is a perceptual one, it is natural to suppose that perception plays at least some important role in our grasp of mental concepts. Theory-theory may also be viewed as containing a conceptual account: grasping mental concepts is a matter of grasping their role in said theory (I call this the *folk-psychological account* in Section 4.2.1). Although the AfC is seldom pitched as a solution to the psychological problem, we will see in Chapter 4 that it draws an important connection between the development of a child's capacity for third-person mental attributions

[44] Robert Pargetter, 'The Scientific Inference to Other Minds', *Australasian Journal of Philosophy* 62, no. 2 (1984): 158.
[45] For an overview of both theory-theory and simulation theory, see Chapter 1 in Jane Suilin Lavelle, *The Social Mind: A Philosophical Introduction* (London: Routledge, 2018).
[46] Hyslop, *Other Minds*, 1995, 71.

and their grasp of mental concepts (and, in turn, the legitimacy of the attributions). Again, a psychological account may not necessarily imply anything about the conceptual problem, but, at least generally, it seems reasonable to suppose that it goes to some lengths in constraining the kinds of ways that we can come to grasp mental concepts.

Lastly, some writers have suggested that rather than go along with the sceptic's schemes (by taking the threat of solipsism seriously), we are better off proceeding with the assumption that we are already justified in believing that other people are conscious and try only to "identify the *basis* of this knowledge."[47] On this attitude, it seems that the POM simply collapses into a psychological problem. In Paul Thagard's words, "this [epistemic POM] is rather silly, since no one doubts that there are other minds. We can dispose of the philosophical problem quickly, then move on to the more interesting and pressing psychological question: given that there are other minds, what can we know about them?"[48]

1.5 Thesis Structure

Each chapter to follow corresponds to the three commonly discussed solutions to the POM: the AA, the ABE, and the AfC. Philosophers have developed different versions of these arguments over time. For the most part, I will stay close to the orthodox versions throughout the thesis, only introducing modifications where necessary. These versions all focus on behaviour as their level of analysis – what separates them is the different ways they use it as evidence. That said, each chapter touches on the relevance of other forms of evidence for their respective framework.

The two first chapters share a similar structure. I begin both with an introduction of the solution as proposed to the POM, followed by its application to the problem of AI consciousness. I then critically examine the solution, delving into the challenges it faces, some of which have been raised against its original application to the POM, and others which arise uniquely in the context of AI consciousness. However, Chapter 3 diverges from Chapter 2 with an additional section dedicated to applying the ABE to three example artificial systems: the imaginary humanoid Replicant, MIT's Cog, and Google's LaMDA.

I argue that there is much that is attractive about the AA and the ABE, at least within the epistemic framework which they operate in (which accepts the asymmetry thesis).

[47] Bayne, *Philosophy of Mind*, 203. See also Section 4.4 in Paul Thagard, *Coherence in Thought and Action* (MIT Press, 2002).
[48] Thagard, *Coherence in Thought and Action*, 102.

Moreover, we will see that it may not be necessary to pit the two arguments against each other, for, as some writers have rightly pointed out, there is much that is compatible between them.[49]

That said, my allegiance ultimately lies with the AfC to be inspected in the fourth and final chapter. Compared to the AA and the ABE, this argument is relatively unexplored, especially in recent decades. In my view, this is a mistake, for (among other things) it offers a response to the conceptual problem to which the proponents of the AA and the ABE have paid little attention. Indeed, it is only within the context of this neglected problem that the force of the AfC can be fully appreciated. I therefore open this chapter by expanding on the conceptual problem, before turning to its consequences for the AA and the ABE. It is here that the two arguments are taken out of their comfort zone, so to speak, and challenged. The AfC is officially developed in Section 4.3. The first subsection here introduces the notion of criteria which plays a central role in the AfC, followed by the development of a Wittgensteinian account of CONSCIOUSNESS. I then apply the argument to the problem of AI consciousness in Section 4.4 which is divided into three further subsections dedicated respectively to the discussion of the notion of criteria with respective to AI consciousness, the application of the argument to the three artificial systems, Replicant, Cog, and LaMDA, and the evaluation of a possible objection from Thomas Nagel.

[49] Andrew Melnyk, 'Inference to the Best Explanation and Other Minds', *Australasian Journal of Philosophy* 72, no. 4 (1994): 482–91; Thagard, *Coherence in Thought and Action*.

Chapter 2: The Argument from Analogy

Recall Jones. He steps on a piece of Lego – inevitably, he yelps aloud and jumps to clutch his foot. You hardly have to think: he is surely in pain. But why do you believe so, and is this belief justified? Perhaps the most natural reply is that his behaviour is more or less exactly how you would respond if you made the same mistake. In other words, you draw an analogy between yourself and Jones.

The argument from analogy (AA) to be explored in this chapter offers exactly this form of reasoning. Given how naturally this reply comes to us, it should be no surprise that it is arguably the first historical solution to the problem of other minds (POM).[50] It was a popular position among philosophers up until the middle of the twentieth century when, facing a number of objections widely considered fatal, it began to fall out of favour.

I begin in Section 2.1 where I explain the AA and apply it to the problem of AI consciousness. Section 2.2 introduces challenges that are thought to face the AA, tailoring them more specifically to its application to the problem of AI consciousness. By the end of the chapter, I hope to have shown that the AA is robust against these challenges and demands more recognition from today's philosophers.

[50] For example, there is evidence of the AA in Saint Augustine's work written no later than AD 428. Augustine, 'De Trinitate', in *Nicene and Post-Nicene Fathers of the Christian Church*, ed. Philip Schaff, vol. 3 (Grand Rapids: Eerdmans, 1974), bks 6, 8, 9.

2.1 What is the Argument from Analogy?

Let us begin with a clear idea of what an analogy is. In its broadest sense, an analogy compares objects (or properties, states of affairs, relations, etc.) for the purpose of highlighting ways in which they are similar. Take, for example, Shrek's claim that ogres are like onions.[51] He is highlighting the fact that ogres have complex layers of personality despite their brutish appearance. An analogy, then, is a tool that helps us infer or grasp something non-obvious about an object. An analogy can be represented in the following general argumentative form:

> *Premise 1*: A and B share properties x, y, and z.
> *Premise 2*: A possesses further property q.
> *Conclusion*: Therefore, we have some reason to believe B has property q as well.

Analogical arguments like these are inductive in nature: they use a limited number of observations to infer a conclusion about unobserved or unobservable things. The confidence an analogical argument confers on its conclusion depends on a variety of factors, such as which objects are being compared, what the extent of their similarity is, and how significant the known shared features are assumed to be.

Analogies are a powerful reasoning tool, and we frequently encounter and employ them in our day-to-day life. But we should be clear about the different kinds of analogies used for different purposes. Many analogies we come across are literary devices intended to help the reader grasp and relate to complex, often abstract, ideas. Importantly, analogies of this kind do not typically draw any literal (as opposed to metaphorical) connection between the compared objects, and the kind of information gained through them is usually aphoristic or emotional in nature. For example: "Life is like a box of chocolate, you never know what you are going to get," "my love for you is as vast as the sea," and "Ogres are like onions."

By contrast, in this chapter I am interested in the kind of analogies intended to establish substantive philosophical or scientific conclusions. For example, physicalists often suggest that since, on their view, consciousness is a physical property, we should expect that it is empirically reducible in a way that water, another physical property, is reducible to its known chemical basis, H_2O. The idea is that since the low-level physical basis of many phenomena have been

[51] Shrek is the titular green ogre protagonist of the film *Shrek*.

revealed by scientific means, we have no reason to suppose that consciousness demands special treatment.

Another philosophical use of analogy is found in animal ethics where a common strategy for defending the moral status of certain non-human animals is to suggest that they share with us a crucial, morally relevant trait: sentience (roughly, the capacity to experience pleasure and pain). Since, it is claimed, humans and many non-human animals share the property of being sentient, and arguably sentience is what bestows moral status to humans, it bestows moral status to those non-human animals as well. Often in the same breath, it is claimed that our ill treatment of them (e.g., factory-farmed animals) is analogous to the treatment of slaves in the past, and therefore criticized as wrong. The plausibility of these particular comparisons aside, we can see that analogical reasoning is a useful tool for revealing something about our world or for justifying a position.

How, then, might analogical-inductive reasoning aid with the POM and the problem of AI consciousness? Let us see how J. S. Mill, a key figure associated with this strategy, puts it:

> I conclude that other human beings have feelings like me, because, first, they have bodies like me, which I know in my own case, to be the antecedent condition of feelings; and because, secondly, they exhibit the acts, and other outward signs, which in my own case I know by experience to be caused by feelings.[52]

Mill's formulation makes it clear that an analogy is being employed.[53] The two things being compared are myself and other people, and the relevant similarity cited is that of the body and behaviour. By 'behaviour', I include facial and verbal expression.

To a first approximation, then, the analogical argument being made is this: since certain conscious experiences of mine typically cause (or, at least, are reliably correlated with) certain behaviour from me, similar behaviour from individuals that look like me provides some reason to believe that they too are conscious. For the proponent of this argument, this reason is good enough to justify this belief.

---

[52] John Stuart Mill, *An Examination of Sir William Hamilton's Philosophy: Volume 9*, vol. 9 (University of Toronto Press, 1979), 191.

[53] That said, Janice Thomas provides compelling contextual evidence that, although analogy is a feature of Mill's argument, his overall argument is one from best explanation (see Section 3.1). Anita Avramides and Anil Gomes make a similar observation. Janice Thomas, 'Mill's Argument for Other Minds', *British Journal for the History of Philosophy* 9, no. 3 (2001): 507–23; Avramides, *Other Minds*, 2000, 168–69; Anil Gomes, 'Skepticism About Other Minds', in *Skepticism: From Antiquity to the Present*, ed. Diego Machuca and Baron Reed (Bloomsbury Academic, 2018), 700–713.

If we adapt this argument to the problem of AI consciousness, we get the following claim: if an AI resembles my body and exhibits behaviour which I know in my own case to be caused by conscious states, then I am justified in believing that it is conscious. This may not mean I have no reason to believe that an AI system that does not resemble me in this way is conscious. So, we should understand, at least for now, the argument as providing a sufficient, rather than necessary, condition for justified belief in an AI's consciousness. Also notice that I have put the argument in general, conditional (if/then) terms since AI systems come in many different shapes and sizes, so to speak. By contrast, the original AA has a clear target: neurotypical human beings. To help evaluate the AA, I want to use Replicant, the imaginary humanoid AI introduced in Chapter 1, as my primary target. The argument will look as follows:

Premise 1: Replicant and I are similar in bodily appearance and behaviour.
Premise 2: I am conscious.
Conclusion: Therefore, I am justified in believing that Replicant is conscious as well.

As it currently stands, however, the AA is incomplete because naturally there will be variables other than behaviour that affect the strength of the inference. First, we should factor in the initial stimuli where possible. After all, what would be the use of behavioural similarity if they were responses to very different stimuli? Think back to Jones. A more complete analogical justification of the belief that he is in pain should include the observation of his stepping on a Lego piece. Stepping on a Lego piece, being bitten by an insect, and eating mouldy bread are all observable stimuli as such. We can also expand this idea into a more general claim about importance of context. If, for example, Jones' behaviour took place on a theatre stage, it would be reasonable to hold that he is not in fact in any pain – he only made it look as if he has stepped on a Lego piece. These considerations are all faithful to the AA.

Next, Mill rightly suggests that bodily modifications – i.e., physiological changes occurring as a result of a stimulus (e.g., swelling or bleeding) – are useful inductive evidence. We should also avoid extrapolating from a small behavioural sample. As Mill points out, one's behaviour should also be consistent to count as reliable evidence. Of course, context-insensitive or irregular behavioural patterns from a fellow human would not compel us to doubt their consciousness (although it may cast some doubt on whether they have an ordinary mind). However, when applying these constraints to non-humans, especially AI systems, constant behavioural anomalies would cast doubt on their status as conscious beings (e.g., if they consistently exhibit pain-behaviour in response to seemingly rewarding activity).

In combining all these variables, we get a more complete picture of the causal train that the AA appeals to:

Input stimuli → Bodily modification → *Conscious experience* → Behavioural output

Let me thus restate the AA: I am justified in believing that an AI system is conscious if, in response to the same kind of stimuli, it exhibits context-sensitive and consistent patterns of behaviour which I know in myself is caused by conscious states.

Before turning to the challenges facing the AA, it is worth recognizing that thinking about the possibility of AI mentality through the analogical lens is not new. Craig Waterman suggests, for example, that we can understand the Turing test on the model of the AA against the mainstream behaviourist interpretation.[54] In claiming that a computer can be considered to think where a human interrogator fails to tell it apart from a human participant in a blind conversation, Alan Turing indeed seems to be relying on an analogy between us and computers based on symbol-generating behaviour. As Waterman describes:

> I could not directly observe a machine thinking any more than I can directly observe the thinking of another human being, but I can observe a machine's production of symbols. A machine that could pass the Turing test would produce symbols like those I know to be instruments of thinking in my own case. Not only would it produce these symbols, but it would be as adept at producing them in conversation as a human being. I know that in my own case such conversion depends on intelligent thought. So, if a machine could pass the Turing test, it would be reasonable to infer by analogy that the machine could think. In fact, it would be arbitrary not to make this inference, since I readily infer that other human beings are intelligent on the basis of behavior that is equivalent to the machine's behavior.[55]

The Turing test has been applied, to varying degrees of faithfulness to Turing's original formulation, to many other domains such as moral status, personhood, emotion, sense of

[54] I should make clear that Waterman thinks Turing's original test is ultimately ambiguous between a (logical) behaviourist and an analogical interpretation. Craig M. Waterman, 'The Turing Test and the Argument from Analogy for Other Minds', *Southwest Philosophy Review* 11, no. 1 (1995): 15–22; A. M. Turing, 'Computing Machinery and Intelligence', *Mind*, 1950, 433–60.
[55] Waterman, 'The Turing Test and the Argument from Analogy for Other Minds', 19.

humour, creativity, and free will in AI systems.[56] If Turing was indeed putting forward an AA, then these attempts (many of them contemporary) to use his test indicates the relevance of the AA as a source for thinking about many epistemic issues relating to AI systems.

2.2 Challenges

The reader should keep in mind throughout this section that I will be evaluating the AA on its own terms – that is, within the conceptual framework which it and the various objections to be discussed operate in. This framework is one that accepts the asymmetry thesis. However, when we come to Chapter 4, I evaluate the AA from a very different perspective – by challenging this very framework.

2.2.1 The unverifiability objection

The first charge against the AA which I will evaluate is that the claim that Replicant is conscious is unverifiable given that we have no direct access to his consciousness. Let us call this the unverifiability objection. Of course, the AA that I have introduced concludes only that we are justified in believing that Replicant is conscious, but no doubt an idealized version of the argument wants to conclude that Replicant is conscious. I will therefore grant that this objection is applicable to the AA. As Carl Wellman puts it:

> The trouble with this inference to other minds is that it is incurably indirect. Since any argument by analogy is non-deductive, no matter how certain one may be of one's premises, there is always the possibility of error in the conclusion. In most cases this uncertainty can be overcome by finding direct evidence for the truth of the conclusion. In the case of other minds, however, this substantiation is impossible. Since one can never be directly aware of the mind of another, there is no way of making sure exactly what is going on in that mind.[57]

[56] Robert Sparrow, 'Can Machines Be People? Reflections on the Turing Triage Test', *Robot Ethics: The Ethical and Social Implications of Robotics*, 2012, 301–15; Manh-Tung Ho, 'What Is a Turing Test for Emotional AI?', *AI & SOCIETY*, 2022; Huma Shah and Kevin Warwick, 'Machine Humour: Examples From Turing Test Experiments', *AI and Society* 32, no. 4 (2017): 553–61; Selmer Bringsjord, Paul Bello, and David Ferrucci, 'Creativity, the Turing Test, and the (Better) Lovelace Test', *Minds and Machines* 11, no. 1 (2001): 3–27; Seth Lloyd, 'A Turing Test for Free Will', *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 370, no. 1971 (2012): 3597–3610.

[57] Carl Wellman, 'Our Criteria for Third-Person Psychological Sentences', *The Journal of Philosophy* 58, no. 11 (1961): 281. Tye also raises this objection in Tye, *Tense Bees and Shell-Shocked Crabs*, 52.

It seems right that, in general, unverifiability is undesirable and should raise suspicion, for it prevents us from spotting false claims from true ones. Whether or not Replicant is conscious, the objection goes, the AA cannot help us verify it.

What should we make of the unverifiability objection? First, one might point out that the fact that a claim is verifiable does not thereby make it credible, for it may turn out to be false. Alec Hyslop and Frank Jackson go further, arguing that verifiability itself confers no epistemic advantage:

> [T]hose who find reassurance in the possibility of directly *verifying* the conclusion of some analogical argument appear to have overlooked the fact that the possibility of directly verifying the conclusion of AA goes hand in hand with the possibility of directly *refuting* the conclusion of AA. . . Therefore, if the impossibility of direct evidence in favor of AA's conclusion counts against the soundness of AA, then, by parity of reasoning, the impossibility of direct evidence against AA's conclusion counts in favor of the soundness of AA. Therefore, the two balance out and leave the soundness of AA untouched.[58]

Hyslop and Jackson seem to be challenging the idea that verifiability is necessarily a virtue – on balance, the value of verifiability is neutral since new evidence may well undermine one's claim.

It is, of course, true that a verifiable claim is always at risk of being undermined by new evidence, but I think the above response misses the point of the unverifiability objection. The motivation for the objection is that the possibility of verification as a whole – i.e., the possibility of confirming *or* falsifying a claim – is valuable. We obviously prefer that our conclusions turn out to be true, but an honest philosopher should be willing to accept whichever conclusion the evidence points to.

Another response to the unverifiability objection is that many (arguably) justified beliefs are unverifiable inductive claims too. Consider the belief that all ravens are black or that the Sun contains helium. Our past observations of ravens provide inductive support for the first belief, even if we are unable search the entire universe for every raven. The fact that samples of helium on Earth has a certain spectrum and that the same spectrum can be observed

[58] A. Hyslop and F. C. Jackson, 'The Analogical Inference to Other Minds', *American Philosophical Quarterly* 9, no. 2 (1972): 169–170.

emitting from the Sun allows us to analogical support the second belief, despite the fact that we cannot travel to the Sun to collect samples of its own helium.

However, the proponent of the unverifiability objection replies, those beliefs are only *contingently* unverifiable. At least in principle, we can scour the universe and we can travel to the Sun to verify those respective beliefs. By contrast, the claim that Replicant is conscious is *logically* unverifiable: no amount of observation, technological advancement, or philosophical scrutiny will give us direct knowledge of his consciousness.

The contingent/logical distinction strikes me as having at least some epistemic significance. In the case of contingently unverifiable conclusions, our curiosities about them can, in principle, be satiated, but for logically unverifiable conclusions, their truth will forever evade us. All things roughly equal, we would certainly favour the former. That said, I think two replies are available. First, it is not obvious (at least not beyond doubt) that the AA's conclusion is indeed logically unverifiable. It certainly is difficult to imagine whether and how it could be verified that Replicant (or any creature other than myself) is conscious, but a lack of imagination is not enough to settle the matter.

The second and, in my eyes, the more crucial reply is this: the fact that the conclusion is logically (or, for that matter, contingently) unverifiable does not entail that the argument itself is a poor one, that it does not adequately support its conclusion.[59] That is, the question of verifiability alone says little about the quality of the inference being made given the available evidence. In short, justification is independent of verifiability. We must not conflate the inability to distinguish between false and true claims with the inability to distinguish between warranted and unwarranted claims. I thus agree with Wellman: "It is no objection to say that I have no way of verifying this statement directly since I can never feel his pain. Of course I cannot feel his pain; that is why I must rely upon an indirect justification for my statement."[60]

The proponent of the unverifiability objection may interject by saying that I am missing the point: the objection is not merely that logically unverifiable conclusions are to be rejected, but that there can be no good argument for them. Could this principle be true? Here is a possible counterexample. Recall (Section 2.1) the analogical argument for the moral status of non-human animals. In short, the argument was that similarity in sentience provides grounds to conclude that non-human animals possess moral status just as we humans do. Now, it is difficult to see whether this conclusion (or any normative conclusion) is logically verifiable.

[59] Wellman, 'Our Criteria for Third-Person Psychological Sentences'.
[60] Wellman, 293. To be precise, in light of what I have said in Section 2.1, what Wellman should have said is: ". . . I can never feel his pain *and know it as such* – as *his* pain."

Yet, the argument is widely regarded by philosophers as compelling. I suspect the proponent of the unverifiability objection is not willing to reject this argument simply for the fact that its conclusion is logically unverifiable. I cannot delve into what makes that argument compelling in the eyes of many, but the lesson should be clear: it seems that there can be good analogical arguments for logically unverifiable conclusions.

2.2.2 The one-case objection

Let us now shift our attention from the AA's conclusion to the argument as a whole: does the AA provide a good justification for believing that certain AI systems are conscious? One reason to answer in the negative is that it makes an inference from a single case and therefore is a very weak inductive argument. Indeed, since I have direct knowledge only of my own consciousness, there is merely one piece of confirming datum from which to infer consciousness in others. I will call this the *one-case objection*.

One response is that one-case induction is warranted where the inference is towards a small number of targets. In the current case, I can be seen as using evidence in my own case to infer that one other being, Replicant, is conscious. In other words, where the inferential ratio, so to speak, is low, one-case induction is warranted. However, a limitation for this response is that the AA should want to conclude more than just that Replicant (or a few other AI systems) is conscious.

That said, I think the above reply can provide a starting point for a more compelling response to the one-case objection. This is that even if it is true that I can use evidence in my own case to justifiably infer the presence of consciousness in just one (or at best a few) AI system, this one system may be *any* system. That is, if the resemblance which holds between me and an AI system like Replicant is sufficient to justify an analogical inference, "then, by parity of reasoning, in *any* unexamined case where the analogy holds, regardless of how many of these there are, we are entitled to make the inference in question."[61]

Another counter to the one-case objection begins by asking why exactly a large sample is useful. The answer is presumably that they together provide greater evidential support. Take the claim that alcohol consumption during pregnancy causes Fetal Alcohol Syndrome (FAS). Clearly, a larger pool of confirming data of alcohol drinking mothers giving birth to children with FAS is beneficial for supporting a link between the two phenomena. But notice that this is a *practical* benefit: more confirming data are useful because we are yet uncertain whether

[61] Hyslop and Jackson, 'The Analogical Inference to Other Minds', 174.

the link is spurious. However, we can imagine an observation of a kind that provides sufficient causal information such that further observation is unnecessary. Suppose, for example, toxicologists discover, through a study of a single pregnant mother, a direct causal mechanism between alcohol consumption during pregnancy and FAS. In such a case, we would be warranted in believing that the finding applies to all pregnant women and their children. It would not *hurt* to look for further confirming evidence, but it will not be strictly necessary. (At any rate, the total number of possible alcohol-consuming mothers is surely too large for several more confirmations to be of any notable statistical benefit.) As Hyslop and Jackson write, "[a]lthough we often demand that an analogical argument be based on more than one instance, nevertheless there are analogical arguments which we take to establish their conclusions with very high probability which proceed from only one examined case. . . AA is of the latter kind."[62] That is, the evidence in my own case sufficiently establishes the causal link (or at least stable correlation) between consciousness and behaviour and thus there is no need to demand further evidence. This assessment seems compelling to me.

2.2.3 The dubious principle objection

Another criticism of the AA is that it makes a logically invalid use of evidence in one's own case. In its standard form, the AA draws upon the introspective evidence of the causal connection going from consciousness to behaviour. However, the inference to which this evidence is applied proceeds in the opposite direction, from behaviour to consciousness – and this commits the fallacy of affirming the consequent. In addition, it is said that the underlying principle that like effects have like causes is a "highly dubious principle."[63] After all, seeing a broken egg does not necessarily mean that it was dropped, even if dropped eggs break (or the fact that gold glitters does not mean that all that glitters is gold).[64] Why should we believe that an AI's behaviour is caused by conscious states because conscious states causes certain behaviour in oneself?[65]

In response to the above worry, Andrew Melnyk suggests that we ought to reason from causes to effects:

[62] Hyslop and Jackson, 174–75.
[63] Hyslop, *Other Minds*, 1995, 37.
[64] Don Locke, 'Just What Is Wrong with the Argument from Analogy?', *Australasian Journal of Philosophy* 51, no. 2 (1973): 153–56.
[65] Of course, this is not to deny that behaviour can cause changes in one's state of consciousness. I am strictly talking about behavioural effects of consciousness.

> Philosophers, I propose, should pay less attention to the *effects* of mental states (i.e. behaviour), and rather more to their *causes* (e.g. in the case of pain, bodily damage). For my own part, I am most confident that someone else is in pain when I see something happen to the other person's body which, I am sure, would cause *me* to feel pain – qualia and all – if it were to happen to me. For instance, if I see someone stub their toe, I don't have to wait for their pain-behaviour to know that they are in pain. I have stubbed my own toe in the past, and I know that toe-stubbing causes pain in the owner of the toe.[66]

Considering the causes of conscious states is no doubt important. However, I see no reason here to pit effects against causes – both seem to me crucial for making the best possible inductive inference. In fact, recall that the AA I have introduced in Section 2.1 incorporates not merely a system's behavioural output, but also (where possible) the type of input, subsequent bodily modifications, as well as the context in which the behaviour occurs. Hence it is false that the AA reasons simply from like effects to like causes. (Or we might put the point slightly differently: when the causal context is taken into account in this way, reasoning from like effects to like causes is *not* dubious.)

Others, however, demand more. Alec Hyslop, for example, suggests that we ought to reason from brain states to mental states.[67] If we modify this proposal to the problem of AI consciousness, the question would be whether an AI system possesses the relevant internal states (e.g., artificial neural states) from which we may infer its consciousness. Indeed, current research into tests of artificial mentality is interested in a wide range of non-behavioural indicators, including architectural, cognitive, and (artificial) neural markers. As James Reggia notes in his 2013 review of the field of artificial consciousness, modern tests artificial consciousness focus on "a system's internal mechanisms, rather than being based solely on

[66] Melnyk, 'Inference to the Best Explanation and Other Minds', 487. Melnyk raises this point in discussing the argument from best explanation (ABE), but the point is nonetheless relevant. John Searle makes a similar claim in John Searle, *The Rediscovery of the Mind* (MIT Press, 1992), 73–74.
[67] Hyslop, 'Other Minds', 2014.

behavioral criteria."[68] Should we, then, favour tests of AI consciousness (and of other mental phenomena) that look "under the hood" for internal markers?[69]

Again, we need not think that behavioural markers and internal markers are in conflict – both seem important to me. However, where behavioural and internal indicators *are* in conflict – that is, where one indicator points in the direction of a 'zombie' system and the other of a conscious system – it is far from obvious which we are to favour and why. For example, it seems to me that I ought to question Jones' status as a conscious being (or at least as an *ordinary* conscious being) if he did not behave as I do in response to the same kinds of stimuli – even if, that is, his brain is healthy and registers the same activity as mine. Similarly, if it turns out that Jones' brain works very differently to mine, I should not, it seems, simply assume that his pain-behaviour, which closely resembles mine, is no longer useful evidence. At the very least, until we have a very good understanding of the relationship between consciousness, its empirical basis, and behaviour, I think it is an open question what we are to say about such scenarios.

Much the same, I suggest, can be said of AI systems: we have little reason to think that we must focus on an AI system's internal mechanisms over its behaviour. In fact, it seems to me that behaviour will play an especially important role since future AI systems are likely to differ from us 'under the hood' at various levels. The material with which their internal mechanisms are realized will probably be silicon-based, not carbon-based like ours. They will be equipped with a shiny CPU instead of a wet brain, and wires will connect and power their body parts rather than nerves and veins. Perhaps most importantly, the nature of an AI system's cognitive architecture will be disanalogous to ours. For example, although many aspects of modern AI design (as well as artificial consciousness research) involve the use of brain-inspired artificial neural networks, architectures based on symbol-manipulation – which follow very different organizational and processing principles compared to the human brain – remains an integral technique as well.[70]

The opponent of the AA may attempt to push back here against the evidential status of behaviour by claiming that the relationship between behaviour and consciousness is too volatile:

---

[68] Reggia, 'The Rise of Machine Consciousness', 115–116. Many contemporary philosophers and scientists agree. Stanislas Dehaene, Hakwan Lau, and Sid Kouider, 'What Is Consciousness, and Could Machines Have It?', 2017, 8; Robert Kirk, *Robots, Zombies and Us: Understanding Consciousness* (New York: Bloomsbury Academic, 2017); Anil Seth, *Being You: A New Science of Consciousness* (New York, New York: Dutton, 2021), chap. 13.

[69] As we proceed, keep in mind that both the ABE and the AfC focus on behaviour too. Some of what follows, with the right adjustments, can be said of these two arguments.

[70] There is, of course, many differences between artificial neural networks and the biological brain too, not the least of which is that the former is a digital network and the latter a physical one.

a mental state (e.g., headache) can cause various types of behaviour as well as no behaviour, and one type of behaviour can be caused by many different conscious states or no conscious state. The AA, however, can mitigate this problem by focusing on the overall behavioural profile of an AI system, rather than narrow items of behaviour.[71] As long as this profile is stable over time (and, as I have said in Section 2.1, context-sensitive), the AA theorist should be satisfied. The evidential status of behaviour used this way, in my view, is secure.

I think it is also worth pointing out that the relationship between internal states and consciousness can be said to be volatile too in various ways. For one thing, our current understanding suggests that the same pattern of neural activity can be correlated with very different kinds of conscious mental activity as well as, more importantly for my purposes, little to no conscious activity.[72] To that extent, we might reason that much the same can be said of an AI system's neural network – there is no easy way to infer its consciousness through their internal states. Further, it may be that consciousness itself can be realized by multiple types of neural states (in the sort of way that a mousetrap can be made with very different materials and designs).[73] If this is right, it would be a mistake to look just for narrow patterns of human-like neural activity. Of course, a future science of consciousness may reveal a more tight-knit relationship, but these considerations provide at least some reason to resist a simplistic bias towards internal markers.

2.2.4 How similar is similar enough?

AI systems even of the distant future will not resemble human behaviour in every way. They presumably will not yawn or shed tears, at least not for the biological reasons we do. Such disanalogies naturally place pressure on the AA. In particular, it raises an important question about sufficient resemblance: to what extent must an AI's behavioural profile resemble ours? This question speaks to both the range of behaviour ("how many aspects?") and its detail ("how closely?").

The question of sufficient resemblance does not seem to trouble the use of AA for inferring consciousness between ordinary human beings, for our behavioural profiles are

[71] Bayne, *Philosophy of Mind*, 205.
[72] Parashkev Nachev, Christopher Kennard, and Masud Husain, 'Functional Role of the Supplementary and Pre-Supplementary Motor Areas', *Nature Reviews Neuroscience* 9, no. 11 (November 2008): 856–69; E. Halgren, 'Mental Phenomena Induced by Stimulation in the Limbic System', *Human Neurobiology* 1, no. 4 (1982): 251–60.
[73] Kenneth Aizawa and Carl Gillett, 'The (Multiple) Realization of Psychological and Other Properties in the Sciences', *Mind & Language* 24, no. 2 (2009): 181–208.

closely matched.[74] However, the task becomes difficult with creatures that differ from us. Sometimes this difficulty only concerns the specific POM – telling what, rather than whether, another creature can consciously experience. For instance, we may wonder whether a chimpanzee's bared-teeth grin indicates friendliness or a sense of threat, but few would doubt that it is conscious. But what about lobsters that bleed blue or relatively basic organisms like earthworms? These cases raise the generic POM: are they conscious at all? The challenge becomes most apparent, of course, when dealing with AI systems. The core challenge as Tim Bayne describes is that "[t]he more an artificial system differs from us, the less appropriate it is to subject it to tests of human consciousness, but the less a test for consciousness is tailored to us, the more difficult it is to validate."[75]

That said, we have some reason to expect AI systems to share some of our behavioural traits. For example, we have an incentive to create robots that act according to goals, react to harmful stimuli, and even converse with us on complex matters. Even where they do not behave exactly as we do, we might reason that they do something functionally similar. They may be said to 'sleep' in the sort of sense laptops do, and they may not share our dietary requirements, but they will nonetheless need some source of energy. Many future AI systems will also, quite convincingly, talk and act at least as if they possess mental states. Companion robots will say things like, "I am *worried* about your smoking habits" and "I *enjoy* reading books for you." Military robots may report things like, "I *see* two unidentified aircrafts to the east," "I *believe* she is bluffing" or "I *sense* fear in their eyes."

It is also worth pointing out any proposed solution to the problem of AI consciousness, including the ABE and the AfC, will face some version of the sufficient resemblance question. The ABE holds that we can justifiably believe that a system is conscious if consciousness best explains its behaviour. Here, it will need some account of what counts as 'best' or 'sufficient' explanation. The AfC holds that a system is conscious if it meets our criteria for consciousness. Here, it will need to explain just how closely it must meet these criteria. In addition, the challenge does not go away by focusing on other levels of description. For example, the AA pitched at the computational level will need to explain just how close a system's computational features must resemble our own. Those who emphasize the role of biological features will need to explain at which point a system is not biological enough – surely an artificial heart is not all

[74] Pargetter argues that the AA cannot, without begging the question, explain why the many differences between oneself and others (e.g., height, hair colour) are irrelevant. This objection seems to me to fail, for introspective evidence, the AA theorist will argue, seems to strongly suggest that those kinds of differences do not matter to the causal chain of a conscious event. Pargetter, 'The Scientific Inference to Other Minds', 160–61.
[75] Tim Bayne, *Philosophy of Mind: An Introduction* (New York: Routledge, 2021), 214.

that relevant, but what about replacing half of the brain's neurons with silicon alternatives?[76] Hence, we should not think that the inability to provide a comprehensive answer to the question of sufficient behavioural resemblance undermines the AA's overall framework.

In fact, perhaps it is altogether a mistake to think that there will be a clear and generalizable threshold at which we become justified in our belief that an AI is conscious. We can understand this idea in at least two ways. One, you might take it to be an inherent feature of consciousness that it has no determinate threshold. That is, you might think that consciousness itself comes in degrees (of, say, complexity or richness). Alternatively, you might take it simply as an epistemic heuristic: since resemblance is a matter of degree, our warrant for attributing consciousness to an AI should also be a matter of degree. This idea follows the principle that the probability we assign to something being the case should be tuned to the amount of evidence we have at hand. Here, finding the appropriate threshold for sufficient resemblance may be less of a matter of *discovery*, as if there is a fact of the matter to be stumbled upon in nature, and more a matter of our *decision* (although this decision need not, of course, be an arbitrary one). I happen to favour the latter reading, for it is unclear to me what exactly it would mean for a creature to be more or less conscious (rather than conscious or unconscious), but the overall idea that we should not think of the community of conscious beings as being defined by rigid borders seems to me sensible.

Still, the fact that there is no clear threshold of sufficient resemblance does not absolve the need to glean justified beliefs from unjustified ones. For example, it seems strange, if not untenable, to hold that we are justified in believing that atoms and rocks are conscious, albeit not as justified as we would be for lobsters and Jones. It seems that our effort is best placed in looking for a middle ground. We should be attuned to the fact that resemblance come in degrees, all the while being alert to any dubious claim to justification. One possible starting point, in my view, is with basic forms of behaviour we associate with avoiding harm on the one hand and seeking reward on the other. It is this capacity that seems to anchor other more specific and readily observable behavioural evidence that the AA theorist should take seriously, such as goal-oriented behaviour and the ability to learn and adapt. Even when we choose to look 'under the hood', it will be helpful to look for mechanisms underlying those behaviour, mechanisms for reinforcing positive stimuli and punishing negative ones. There is certain to be variance in what this precisely means for different kinds of AI systems, but it seems to me a crucial consideration.

[76] Tye, *Tense Bees and Shell-Shocked Crabs*, 182–89.

2.3 Conclusion

Let us revisit the key moments in this chapter. We first began by laying down a clear understanding of what an analogy is and narrowed our interest to its use in philosophical arguments to justify beliefs in unobserved or unobservable phenomena. The AA's reply to the problem of AI consciousness is that where an AI system exhibits consistent, input- and context-sensitive patterns of behaviour which I know in my own case (and, more broadly, in human cases) to be caused by or reliably correlated with states of consciousness, I am justified in believing that it is conscious.

We then turned to the challenges facing this argument. The first challenge – the unverifiability objection – attacked the AA's conclusion, whereas the one-case objection and the dubious principle objection attacked the overall strength of its inference. My view was that the AA has robust replies available to these. We then considered a more practical hurdle regarding what constitutes sufficient resemblance. Although it remains unclear how the AA is to fully clear this hurdle, I have emphasized that it is not a problem unique to the AA (or, for that matter, the problem of AI consciousness) as well as suggest ways one might begin to navigate through it.

Although I aim to show in Section 4.2 that there are conceptual reasons to reject the AA, it has otherwise all too easily been dismissed by philosophers. In my view, we should not be so quick to deny its place as a helpful framework for tackling the problem of AI consciousness.

Chapter 3: The Argument from Best Explanation

Recall Jones and his unfortunate encounter with a piece of Lego. He yelps aloud and jumps to clutch his foot. Perhaps you laugh, or pity him, but you believe – perhaps even know – that he is in pain. What might justify this confidence? According to the *argument from best explanation* (ABE), we are justified because pain best explains his behaviour.[77]

The ABE shares with the argument from analogy (AA) the key assumption that there is a deep asymmetry between first- and third-person knowledge of consciousness – one's knowledge of other people's consciousness, in contrast to the self-knowledge of consciousness, can only ever be indirect. In turn, like the AA, the ABE suggests that some form of non-deductive inference is required. But where the AA used analogical-inductive inference – using one's self-knowledge of pain and its connection to behaviour as inductive evidence for generalizing about others – the ABE uses *abductive* inference. As we will see in Section 3.2, proponents of ABE have claimed that this difference allows it to address or altogether bypass some of the key issues facing the AA.[78]

Indeed, although the AA had long been the traditional solution to the POM, the pendulum has swung in the ABE's favour in recent decades.[79] It has proven attractive to

[77] The ABE is often called the *inference to the best explanation* (IBE). In my view, it is best reserving this term for describing the more general principle of inference sometimes called abductive (or theoretical, hypothetical) inference, and using 'the argument from best explanation' to refer to the application of abductive inference specifically to the POM and the problem of AI consciousness.
[78] Pargetter, 'The Scientific Inference to Other Minds'; Paul M. Churchland, *Matter and Consciousness*, The MIT Press (The MIT Press, 2013), chap. 4; Tye, *Tense Bees and Shell-Shocked Crabs*, chap. 4.
[79] Alec Hyslop suggests that the ABE "wins hands down" in popularity, not just against other arguments, but also against the view that there is no plausible solution. Hyslop, *Other Minds*, 1995, 29.

philosophers of varying metaphysical orientation. For example, David Chalmers, a 'naturalistic dualist' in his words, thinks that the ABE "is as good a solution to the problem of other minds as we are going to get."[80] The functionalist phase of Hilary Putnam seems to agree that this is the right way to go, as does Jerry Fodor, whose view on consciousness, as I understand it, is a form of non-reductive physicalism. [81] Lastly, Paul Churchland, a reductionist about consciousness, also endorses the ABE, adding that it can be useful even in cases of machines.[82]

In this chapter, I will carefully trace the steps in this popular argument and consider whether it can help answer the problem of AI consciousness. I begin with an introduction to the ABE framework (Section 3.1), before critically examining it (Section 3.2). Having defended the general ABE formula, section 3.3 examines what it looks like in practice when applied to three artificial systems: Replicant, Cog, and LaMDA.

3.1 The Argument from Best Explanation Explained

Much like inductive inference, abductive inference has an important place in our toolbox for reasoning about the world. This is true whether you are a scientist, a philosopher, a detective, or just an ordinary person. Let me begin with an example to illustrate just what this form of reasoning involves.

Suppose Susan, whom you have just met, tells you she owns a house in downtown Manhattan. Well-aware of the sky-high housing prices around the area, you immediately think to yourself, "she is doing well for herself." Here, you are probably reasoning abductively: you feel that the *best explanation* of her ability to afford such a house is that she has a well-paying job. The reasoning may be a sub-personal (automatic, implicit) one rather than a conscious, effortful one, but it is abductive in nature either way. Of course, there are other possible explanations. The house may have been inherited, or perhaps there are a handful of affordable houses in downtown Manhattan if you search hard enough. Alternatively, she could have simply made a terrible financial decision. Now, strictly speaking, these explanations are not mutually exclusive. For instance, Susan may well be earning a good wage *and* have inherited the house. Nonetheless, they do tend to compete to be the *best* explanation and negatively

[80] David J. Chalmers, *The Conscious Mind: In Search of a Fundamental Theory* (Oxford University Press, 1996), 246.
[81] Jerry A. Fodor, *Psychosemantics: The Problem of Meaning in the Philosophy of Mind* (MIT Press, 1987); Hilary Putnam, ed., 'Other Minds', in *Philosophical Papers: Volume 2: Mind, Language and Reality*, vol. 2 (Cambridge: Cambridge University Press, 1975), 348–49. Note that Fodor here is talking about beliefs and desires, but the broader context of this quote, which involves his cat, mentions pain-behaviour and various other behaviour which suggests he would say much the same about conscious states.
[82] Paul M. Churchland, *Matter and Consciousness*, The MIT Press (The MIT Press, 2013): 118.

constrain each other to some extent. In this case, given what we roughly know about the world, it strikes us as the single most likely possibility that Susan earns a good wage.

What makes a particular explanation the best one? The answer is not always clear-cut, but some of the usual factors – theoretical virtues – to consider are predictive power, simplicity, and congruence with our background understanding of the world.[83] These three virtues are what I will be focusing on in Section 3.3.

Predictive power refers to the likelihood of an explanation to correctly guess a system's future states (e.g., behaviour) or its other properties. In Susan's case, the 'good-wage' explanation may predict that she frequently goes furniture shopping and has a nice car – intuitively, these seem likely. Predictive power is often the most important hallmark of a good explanation, for it speaks to its ability to account for a system's causal profile – what it is caused by and what it in turn causes.

Simplicity is typically understood to concern the number of entities an explanation posits.[84] On this characterization, the simplicity condition states that, all things roughly equal, we should favour the explanation with less posits. However, as I will show in Section 3.2, an explanation with the least posits is not always the simplest option. For this reason, I want to characterize simplicity in a broader way: how much mental gymnastics (or stretch of the imagination) is involved in an explanation. For example, it would require less mental gymnastics to believe that Susan earns a good wage, than that she is, say, married to a Saudi prince.

Finally, the congruence condition concerns how well the proposed explanation fits into our pre-existing network of beliefs about the world. If an explanation goes radically against the grain of our ordinary assumptions and understanding, we have some reason to treat it with suspicion. Importantly, this condition does not strictly require that those beliefs are themselves correct – rather, the emphasis is just that we should not, all things roughly equal, deviate from the epistemic status quo.

Of course, what we take to be the best explanation may turn out to be false. Abductive arguments are not deductive arguments – their conclusions do not *logically* follow from the premises. Nonetheless, the ABE proceeds under the widely held assumption that it is rational

[83] Samuel Schindler, 'Theoretical Virtues: Do Scientists Think What Philosophers Think They Ought to Think?', *Philosophy of Science* 89, no. 3 (2022): 542–64; Peter Lipton, *Inference to the Best Explanation*, 2nd ed. (London: Routledge, 2004).

[84] This is how the ABE literature often frames it. Nathan Stemmer, 'The Hypothesis of Other Minds: Is It the Best Explanation?', *Philosophical Studies* 51, no. 1 (1987): 109–21; Melnyk, 'Inference to the Best Explanation and Other Minds'.

to believe in the hypothesis that provides the best explanation of our observations given the state of knowledge at the time.

Let me now turn to how abductive reasoning is used in science. This will help us fill out the further details of this form of reasoning that are relevant to the ABE. A good example concerns the electron. Searching for an explanation for why cathode rays (for our purposes, picturing a light beam will do) were being deflected by negatively charged plates in his experiments, English physicist J. J. Thomson remarked that he can "see no escape from the conclusion that they are charges of negative electricity carried by particles of matter."[85] In other words, he believed the best explanation for this peculiar observation is the presence of negatively charged particles. He went onto call these particles, *electrons*. To rule out alternative explanations, Thomson made sure to repeat the experiment with adjustments in his apparatus and control other variables.

Now, even today's most powerful microscopes are yet to allow us to directly observe electrons – they remain, strictly speaking, a theoretical posit. Despite this fact, their immense explanatory power has given them an indispensable place in our best scientific theories (e.g., of chemical bonding, magnetism, and conductivity). In fact, most scientists today seldom treat electrons as a mere theoretical posit, and our terminology around them has become vividly realist – so much so that Thomson is almost universally credited as having *discovered* the electron.

According to Paul Churchland, it is precisely because of this sort of success that abductive reasoning has found in the natural sciences that philosophers have started to take it seriously as a tool for tackling the POM:

> The problem of other minds was first formulated at a time when our grasp of the nature of theoretical justification was still rather primitive. Until fairly recently, almost everybody believed that a general law could be justified only by an inductive generalization from a suitable number of observed instances. . . This idea might have been adequate for laws connecting observable things and properties with one another, but modern science is full of laws governing the behavior of *unobservable* things and properties. . . Plainly, laws concerning unobservables must enjoy some other form of empirical justification, if they are to be justified at all.[86]

[85] J. J. Thomson cited in Peter Achinstein, *The Book of Evidence* (Oxford University Press, 2001), 17.
[86] Churchland, *Matter and Consciousness*, 116–117.

It is unclear just how recent Churchland takes this shift to have taken place. It is worth noting here that J. S. Mill's argument for other minds, formulated all the way back in 1865 and commonly believed to be an AA, is understood by some commentators to be abductive at its core. As Janice Thomas, one such commentator, points out, Mill likened his argument to Isaac Newton's *universal gravitation hypothesis* which proposed that both the falling of objects on Earth (e.g., an apple from its tree) and astronomical movements (e.g., planetary orbits) are governed by the same force – namely, gravity.[87] Newton's brilliance, Mill observed, lay not in trying to show that there could be no other forces acting upon these phenomena, but in simply positing no more forces than necessary to adequately explain them. That is, that universal gravitation is the best (in particular, the simplest, most unified) explanation. Similarly, Mill held that, given similar sensory input and similar behavioural output, there is little need to posit the absence of mental states in other people. In drawing this comparison, Mill indeed seems to have given us the groundwork for the ABE.

That said, Churchland would be right that systematic debate around the ABE is indeed a relatively recent development (of the last fifty years or so), and its proponents often do appeal to the fact that it builds on the success abductive reasoning has found in the scientific domain.[88] Let us, then, see what this argument looks like.

According to the ABE, we are justified in believing that other people are conscious since this belief – call it the *consciousness-hypothesis* – best explains their behaviour. Take our friend Jones. According to the ABE, the pain-hypothesis best explains his behavioural response to stepping on a Lego piece and therefore we are justified in believing that he is in pain (and, *a fortiori*, is conscious). As Robert Pargetter explains, "On this account, my reason for believing in the existence of other minds similar to my own is in all important respects the same as the scientific realist's reason for believing in the existence of sub-atomic particles."[89] The shared reason being, of course, an *explanatory* one. Just as Thomson did not require certainty about the existence of a single electron to hypothesize the existence of electrons more generally (i.e., use inductive reasoning), the fact that *I* am conscious, the thought runs, does not figure in the inference that others are conscious. All that is required is that the relevant hypothesis best

[87] Thomas, 'Mill's Argument for Other Minds'; Mill, *An Examination of Sir William Hamilton's Philosophy*, 9:190–91.
[88] See, for example, Pargetter, 'The Scientific Inference to Other Minds'; Charles S. Chihara and Jerry A. Fodor, 'Operationalism and Ordinary Language: A Critique of Wittgenstein', *American Philosophical Quarterly* 2, no. 4 (1965): 281–95.
[89] Pargetter, 'The Scientific Inference to Other Minds', 159.

explains our observations. Under this view, then, the consciousness of other people is, strictly speaking, a theoretical posit.

We need to tread carefully in applying the ABE to AI systems, for whereas its application to the standard POM has a clear, uniform target – neurotypical human beings – AI systems come in many different forms. Whether we are warranted in believing that an AI system is conscious according to the ABE therefore depends in part on what kind of system we are dealing with. I will express the ABE first in conditional terms (as I did with the AA): if the consciousness-hypothesis best explains an AI's behaviour, then we are justified in believing that it is conscious. Next, let us give body to this conditional claim using the imaginary humanoid Replicant as our target system.

In order to examine the plausibility of the consciousness-hypothesis for Replicant (and AI systems in general), we should be clear about the alternative hypotheses it is competing against. Here, given that my focus lies with whether an AI system is conscious (the generic POM), rather than what the character or content of its conscious experiences is (the specific POM), alternative hypotheses that propose that Replicant is conscious, but in some very different way, will not be of interest to me, at least not in any way distinct from the consciousness-hypothesis. A classic example here would be the *inversion-hypothesis*: pain and pleasure are *inverted* for Replicant, and this inversion is systematic in a way that it does not show up in his behaviour.[90] The inversion-hypothesis is simply a version of the consciousness-hypothesis (call it, if you will, the *inverted-consciousness-hypothesis*).

Instead, the alternative hypothesis that I focus on is that Replicant is not conscious. Although he talks and acts very much as we do, he lacks the subjective perspective from which he experiences the world – there is no light inside him, so to speak. His behaviour is to be explained by non-conscious physical states. I will call this the *zombie-hypothesis*, for it proposes that although Replicant, sort of like a zombie, resembles us in a great number of ways, lacks consciousness.

Before I turn to the next section, it is worth distinguishing the ABE from a well-known framework of mental attribution: Daniel Dennett's *intentional systems theory*. According to this view, anything – other humans, animals, and even *robots and computers* – whose behaviour is "usefully and voluminously predictable from the intentional stance is, by

[90] Ned Block, 'Troubles with Functionalism', *Minnesota Studies in the Philosophy of Science* 9 (1978): 261–325.

definition, an *intentional system*."[91] His focus here lies with intentional states (e.g., beliefs, desires), but I will, as some writers have done, extend the theory to include conscious (or phenomenal) states such as pain.[92] Call this the *phenomenal systems theory*.

The ABE and the phenomenal systems theory share the appreciation of predictive power with respect to behaviour as a basis for attributing consciousness to others, including artificial systems. However, the latter, by offering a *definition* of a conscious system, is committed to the view that there is *nothing more* to being a conscious system than being a system for which our behaviour-based interpretation of them as such is predictively useful. The ABE, by contrast, does not commit to such a strong view. Instead, it need only be committed to the view that where a phenomenal stance towards a system's behaviour is predictively (and, more broadly, explanatorily) useful, we can non-deductively infer their consciousness.

3.2 Evaluating the ABE

The ABE is often touted as an improvement on the AA, avoiding the objections that many philosophers have felt to be the latter's downfall. If you recall, my view was that the AA can in fact withstand these objections. Nonetheless, let us revisit them and see how the ABE might respond to them. (As with the AA in Section 2.2, I evaluate the ABE here within its own conceptual framework. I challenge this framework, however, in Section 4.2.)

First, the unverifiability objection may be raised against the ABE: whereas the financial source of Susan's Manhattan home is straightforwardly verifiable (say, by, befriending her or somehow acquiring her bank statement), and the existence of electrons seems verifiable at least in principle (say, through advanced quantum microscopy), the claim that Replicant is conscious is logically unverifiable. Here, the AA's replies are all equally available to the ABE, but let me revisit just one: the fact that this claim may be logically unverifiable does not itself undermine the quality of the argument being made in support of it. In fact, the unverifiability objection merely highlights the very need for an indirect inference to Replicant's consciousness.

[91] Daniel Dennett, 'Intentional Systems Theory', in *The Oxford Handbook of Philosophy of Mind*, ed. Ansgar Beckermann, Brian P. McLaughlin, and Sven Walter, 1st ed. (Oxford University Press, 2009), 339.
[92] Philip Robbins and Anthony I. Jack, 'The Phenomenal Stance', *Philosophical Studies* 127, no. 1 (1 January 2006): 59–85. Note that I am not attributing this position to Dennett. As some writers suggest, the intentional systems theory is "completely independent of Dennett's views on consciousness." Marc V. P. Slors, 'Intentional Systems Theory, Mental Causation and Empathic Resonance', *Erkenntnis* 67, no. 2 (2007): 329.

What about the one-case objection? It seems the ABE has a unique response available here: since it does not use inductive inference, the objection is altogether inapplicable.[93] That is, since the ABE is an *abductive* argument which proceeds solely on explanatory grounds, "appeal to one's own case plays no greater evidential role in justifying belief in other minds than it does in justifying belief in electrons, i.e. no role at all."[94]

According to its proponents, not only does the ABE bypass the one-case objection, but it is also superior to the AA in accounting for the behaviour of creatures other than neurotypical humans:

> It is not surprising then that the [ABE] not only gives us a good account of our belief in other minds, but also shows us why we are able to make other related inferences justifiably. I infer about some other people that they have abnormal minds, I make inferences about the mental experiences of animals, and I can imagine circumstances where I might infer that alien beings have mental lives. In such cases, the [AA] either is hopelessly weak or is just inapplicable. The very marked differences in behaviour patterns, to the point of very little similarity, would prevent all but the weakest of analogical inferences. But the [ABE] naturally accounts for such inferences because of the explanatory power of the respective hypotheses.[95]

> [O]ne's justification here need owe nothing at all to one's introspective examination of one's *own* case. . . Conceivably one's own case might even differ from that of others. But this need not affect one's theoretical access to their mental states, however different they might be from one's own. One would simply use a different psychological theory to understand their behavior, a theory different from the one that comprehends one's own inner life and outer behavior. . . [We are] justified in ascribing conscious intelligence to other animals, and even to machines, if the explanatory strategy [is] successful in their case also. [96]

[93] There is room for confusion here because some philosophers have called the ABE an *inductive* argument in the outdated sense that includes abduction. See, for example, Hyslop, *Other Minds*, 1995, 31; Pargetter, 'The Scientific Inference to Other Minds', 159.
[94] Melnyk, 'Inference to the Best Explanation and Other Minds', 482.
[95] Pargetter, 'The Scientific Inference to Other Minds', 161–62.
[96] Churchland, *Matter and Consciousness*, 118.

The ABE's proponents will argue, then, that their argument is better equipped than the AA for answering the problem of AI consciousness. Before continuing my evaluation of the ABE, I want to pause to challenge this treatment of the AA.

For one, Pargetter appears to cherry-pick the differences whilst overlooking the similarities. The cases he mentions – in his words, abnormal people, non-human animals, and aliens – are certainly different from neurotypical human beings in many ways, but they are also similar in numerous other ways. Individuals with motor and cognitive disability (e.g., akinetic mutism), many non-human animals, and surely certain aliens, at least those typically portrayed in popular culture, do behave in complex ways recognizably similar to neurotypical humans. Such behaviours include avoiding bodily damage, undergoing sleep/wake (or 'rest/active') cycles, and communicating with others. Since these behaviours are closely associated with conscious experiences in my own case, I seem to have at least some analogical evidence that these other creatures too have conscious experiences. At the very least, the AA is certainly not 'hopelessly weak' or 'just inapplicable' as Pargetter claims.

What about rudimentary animals such as jellyfishes, oysters, and earthworms, or patients with unresponsive wakefulness syndrome (commonly known as the persistent vegetative state)? In these cases, there are challenges in drawing a behavioural analogy not so much because the target creatures behave very differently to neurotypical humans, but because they lack complex behaviour altogether. However, as I highlighted in Chapter 2, the AA is free to draw analogies at other levels of description, such as neurophysiology, cognitive architecture, and evolutionary proximity (what these inferences yield, of course, is a separate issue). It is also worth pointing out that identifying consciousness in these cases is a difficult task regardless of the framework used, including the ABE – e.g., it is not obvious that consciousness best explains worm behaviour or physiology. Lastly, it is not quite true that there is no useful analogical-behavioural evidence of consciousness in some such cases. Notably, even oysters and worms respond to noxious stimuli – and I see no reason why this is fundamentally different from human avoidance behaviour.

Now, if Pargetter (or Churchland) has in mind creatures whose behaviour (and other levels of description) differs from ours so much so that the AA is indeed of little use, then it is unclear to me how the ABE would fare much better. The assertion that the ABE can simply develop alternative psychological theories to explain their behaviour seems merely speculative, for we do not know what those theories might look like, let alone whether any will be compelling. Just how are we supposed to justifiably theorize consciousness in creatures whose behaviour evades *any* useful comparison to the behavioural evidence of human consciousness?

Even if the ABE has some theoretical advantage over the AA on this matter, I see little reason to give this advantage the weight that Pargetter (and Churchland) does.

Let us turn our attention back to the claim that appeal to one's own consciousness (I will call this *self-appeal*) is not an essential component of the ABE. One might think to push back against this claim by invoking the prima facie importance of self-appeal for our ability to (1) grasp the concept of consciousness, and (2) describe the consciousness-hypothesis as one among the possible explanations of another's behaviour. The first concern stems from the intuitive idea that grasping the concept CONSCIOUSNESS requires some form of introspective attention to one's own consciousness. If this is right, some form of self-appeal seems necessary just to think about consciousness, not to mention formulate and support the consciousness-hypothesis. The second concern brings into question the possibility of describing the consciousness-hypothesis without the certainty of the existence of consciousness (and its causal link to behaviour) that can arguably only be found introspectively.

Fortunately for the ABE, both worries are founded on a misunderstanding: neither grasping the concept CONSCIOUSNESS nor describing the consciousness-hypothesis – even if it requires introspection – involves or implies *supporting* the consciousness-hypothesis. In other words, the use of introspective facts for such purposes does not constitute *appealing* to those facts to support the consciousness-hypothesis. As Pargetter writes, "Notice that this description [of the hypothesis] by analogy is not an analogical inference. The analogy in no way argues for the truth of the hypothesis."[97]

A genuine worry for the ABE, however, is whether it can support the consciousness-hypothesis over the zombie-hypothesis without self-appeal – i.e., on purely explanatory grounds. This question, in my view, presents the ABE with a major obstacle stemming from our limited understanding of the explanatory role of consciousness.[98] A key requirement for the successful use of abductive reasoning is that the explanatory credentials of the relevant posit are clear and robust. Recall Susan from my earlier example. The explanatory role of a well-paying job is obvious: paying for her Manhattan home. Likewise, electrons play a powerful explanatory role in many scientific theories.

By contrast, we lack a clear, systematic understanding of the explanatory role of consciousness. The extent to which we can evaluate, for example, the predictive power of the consciousness-hypothesis seems to rest crucially on the assumptions we make about the causal

[97] Pargetter, 'The Scientific Inference to Other Minds', 162.
[98] Melnyk, 'Inference to the Best Explanation and Other Minds'; Chapters 2 and 3 in Hyslop, *Other Minds*, 1995; Hyslop, 'Other Minds as Theoretical Entities'.

(or even just correlative) profile of consciousness since only where it makes causal differences can it be said to make behavioural and predictive differences.[99] However, there is currently no widely agreed upon account of the causal profile of consciousness, and we have little clue as to when, if at all, it will become available. Moreover, even if such an account does arrive, what will it say about the nature and extent of that role, particularly with respect to behaviour? We simply do not know.

Why is there such uncertainty and disagreement? As I have briefly touched on in Section 1.1, the central challenge for the study of consciousness is that the way we most naturally think of conscious experiences is in terms of how they *feel* from a subjective perspective – i.e., its phenomenal character – rather than in terms of its causal or functional profile. Take pain, for example. We certainly speak of pain having a causal profile, but we seldom identify pain *as* its causal profile – we identify it as something that one *feels*. This is precisely why the zombie-hypothesis (and the POM more generally) has troubled philosophers: it seems conceivable that a system could exhibit consciousness-behaviour (as well as possess other physical features associated with consciousness) without being conscious.

The range of views one might take on this matter is very broad. On one extreme end, epiphenomenalists deny that consciousness has any causal power. Indeed, discussions on the ABE frequently emphasize that a non-epiphenomenalist view of consciousness is essential for the ABE to work (and, more generally, for the POM to be solvable).[100] On another extreme end, panpsychists believe that consciousness is a fundamental and ubiquitous feature of our universe. On this view, it is not even clear what it would mean for consciousness to have an explanatory role with respect to behaviour.[101] Some views, typically physicalist ones, provide us with some roadmaps for advancing our understanding of the explanatory status of consciousness, but there is little guarantee that the promises will be delivered (and little agreement on which view will come out on top). Therefore, there are challenges to critically evaluating the overall explanatory power of the consciousness-hypothesis and, in turn, to validating the idea that the ABE succeeds on solely explanatory grounds.

Let me borrow Andrew Melnyk's thought experiment to help drive home the current challenge: would a powerful computer (suppose it is non-conscious) equipped with the capacity

[99] More specifically, consciousness must make causal differences that non-conscious states cannot.
[100] That is, as far as behaviour is concerned. Epiphenomenalists may well argue that certain brain states are best explained by consciousness, for they do not deny that consciousness itself can be caused. Pargetter, 'The Scientific Inference to Other Minds', 162–63; Hyslop, *Other Minds*, 1995, 31–32.
[101] Of course, panpsychists may have an answer here (as may other positions), but my main point is simply that there is no consensus.

for abductive reasoning reliably posit consciousness to explain the behaviour of human beings?[102] Melnyk himself thinks that the computer will surely opt for the zombie-hypothesis: "why on earth would it even *occur* to this computer to formulate the [consciousness]-hypothesis?"[103] But even a neutral answer here is enough to challenge the ABE: on our current understanding, there is no clear reason to think the computer would choose to posit consciousness over and beyond purely physical states for human beings, let alone AI systems. Again, take pain. The computer might think to analyze Jones' behaviour by attributing the behavioural disposition for bodily preservation, but it is unclear why it would attribute the feeling of pain.

Now, is this limitation *fatal* for the ABE? There is some reason to answer in the negative. First, it seems reasonable to appeal to our common-sense (or 'folk-psychological') assumptions about the explanatory role of consciousness. For instance, in everyday contexts, we assume that conscious experiences make a causal difference to behaviour. Children in pain cry, as do adults when moved by a sad story. We tend to laugh when tickled and scratch our itches. This is not, I should clarify, to simply help ourselves to the assumption that others are conscious. Rather, it is appealing to the apparent explanatory success of the assumption that others are conscious in everyday settings. Perhaps our common-sense explanations are not trustworthy by the highest of scientific or philosophical standards, but they sure are integral aspects of our daily lives. Deferring to and incorporating these ordinary practices and intuitions into the explanatory equation where systematic understanding is lacking seems sensible.

Second, when we take a step back from all the meticulous metaphysical debates about consciousness, one notices that many scientists and philosophers often do assign explanatory roles to consciousness, not merely in relation to behaviour but to various other kinds of phenomena and processes. Many believe that consciousness has an evolutionary (i.e., fitness-enhancing and adaptive) function[104]; that it is tied to our capacity for various forms of attention[105]; that it enables flexible cognitive and behavioural control[106]; and that it is integral to generating our sense of self and free will.[107] Of course, not all of such roles (e.g., evolutionary function) may be applicable to AI systems, but insofar as many are, we have some

[102] Melnyk, 'Inference to the Best Explanation and Other Minds', 485.
[103] Melnyk, 485.
[104] Brian Earl, 'The Biological Function of Consciousness', *Frontiers in Psychology* 5 (5 August 2014): 697.
[105] Christopher Mole, 'Consciousness and Attention', in *The Oxford Handbook of the Philosophy of Consciousness*, ed. Uriah Kriegel (Oxford University Press, 2020), 0.
[106] Chapter 10 in Bernard J. Baars, *A Cognitive Theory of Consciousness* (Cambridge University Press, 1988).
[107] Thomas Metzinger, *Being No One: The Self-Model Theory of Subjectivity* (MIT Press, 2003); Daniel M. Wegner, *The Illusion of Conscious Will* (Cambridge, Massachusetts: MIT Press, 2002).

grounds to think that explanatory considerations can support the consciousness-hypothesis for them.

All that said, the appeal to common-sense and scientific/philosophical beliefs about the explanatory role of consciousness require a crucial caveat: they must not be grounded on introspective evidence, for if they are, then it seems the ABE would not be proceeding on purely explanatory considerations. For example, if the pain-hypothesis that we naturally accept of Jones' behaviour implicitly relies on facts derived from each of our own pain-experiences, then that hypothesis cannot be said to succeed on explanatory grounds alone. Likewise, if the scientist's claim that consciousness explains our capacity for adaptation and attention appeals at any point to certain introspective evidence, then, once again, any inference that others are conscious based on such capacities would be proceeding from less than purely explanatory consideration. Unless this caveat is met, the appeal to common-sense and scientific/philosophical beliefs are merely pushing the concern over the unclear explanatory role of consciousness one step back rather than genuinely addressing it.

A very different response to the above explanatory concerns is to simply embrace the evidence in one's own case and give up the idea that the ABE must proceed on purely explanatory grounds.[108] One version of this response is as follows:

1) To explain Jones' (or Replicant's) behavioural response to stepping on a Lego, I form two competing hypotheses: the pain-hypothesis and the zombie-hypothesis.
2) On grounds of his behaviour alone, it is unclear which hypothesis is superior, for the explanatory role of pain is unclear.
3) But introspection shows that I experience pain and these experiences are reliably correlated to certain behaviour.
4) Since Jones closely resembles me and my behaviour, the result of this introspection provides evidence for favouring the consciousness-hypothesis for Jones.

The resulting argument, then, is no longer a purely abductive argument – it no longer operates on just explanatory grounds. Instead, it can be seen as a *hybrid* argument that combines the core elements of the AA and the ABE.

[108] Melnyk, 'Inference to the Best Explanation and Other Minds'; Hyslop, *Other Minds*, 1995. Interestingly, Pargetter seems at times to endorse this strategy, contradicting his own claim that the strength of ABE rests solely on explanatory grounds. For example, he writes: "Although there are other possible explanations, [the pain-hypothesis] appears in the light of evidence from my own case to be a very good explanation." Pargetter, 'The Scientific Inference to Other Minds', 158.

The self-appeal being made here is not that the consciousness-hypothesis explains my own behaviour well and therefore that I am a confirming instance of the consciousness-hypothesis. Since the ABE, by accepting the asymmetry thesis, holds that I simply and directly *know* that I am conscious, it is unclear just how the consciousness-hypothesis can be said to succeed in explaining my behaviour. Rather, the idea is that in comparing competing hypotheses, one can introduce other facts that affect their plausibility, although the hypotheses do not explain those facts. In the hybrid ABE, this other fact is my introspective knowledge of consciousness and its relation to my behaviour.

To clarify this strategy, consider the following case. If a twin child develops a rash today, I may think to explain it by an allergy-hypothesis. The fact that last week the other twin had a doctor-diagnosed allergic reaction makes that hypothesis more plausible, but the hypothesis that *this* twin is having an allergic reaction right now cannot, it seems to me, explain the fact that the *other* twin had an allergic reaction last week.[109] The rough comparison to electrons will be this: if I were to directly observe an electron (and it does the things that the electron-hypothesis predicts), this would reinforce the likelihood that the electron-hypothesis is true. But, again, this is not because the electron-hypothesis somehow explains this observation – it is because, in short, I now know that electrons exist.

Now, the trade-off of being able to support the consciousness-hypothesis through the hybrid approach is that we are confronted with a version of the one-case objection: one's own case gives exactly one piece of evidence in favour of the consciousness-hypothesis. The advantage this single piece of datum confers to the consciousness-hypothesis, one might argue, is surely too weak to really give the consciousness-hypothesis any confident nod over the zombie-hypothesis. The hybrid ABE, however, can make use of the AA's successful response to the one-case objection I discussed in Section 2.2.1. In summary, those who wish to use the ABE for the problem of AI consciousness have a choice: to put their chips into a non-hybrid ABE and trust that consciousness will turn out to have explanatory roles, or to go with the hybrid ABE. My view is that, between these two, one is safer with the latter.

[109] I thank Andrew Melnyk for both his clarificatory comments and introducing me to this example via email correspondence. Note, however, that this way of understanding the hybrid ABE is different from that which he discusses in Melnyk, 'Inference to the Best Explanation and Other Minds'.

3.3 AI Consciousness and the ABE

Let us now explore, with the help of the three theoretical virtues I introduced earlier – predictive power, simplicity, and congruence – what the (hybrid) ABE might say about some example artificial systems. Much of the discussion to follow is neutral between the non-hybrid and the hybrid ABE, but it may help to take the stance of the latter. I will first develop the case of Replicant in further detail, and then turn to two relatively basic but real-life examples, Cog and LaMDA.

3.3.1 Replicant

First, since Replicant closely resembles us human beings in appearance and behaviour, it seems he will closely match whichever predictive success the consciousness-hypothesis yields for us. Of course, the materials with which he and his internal mechanisms are realized radically differ from ours, and certain aspects of his behaviour (e.g., diet, film choices) will be rather strange. However, even if material differences may be *sometimes* relevant for inferring consciousness (i.e., weaken the inference), I see little reason why they should invalidate the overall predictive success of the consciousness-hypothesis in Replicant's case.[110] Consider, for example, the various artificial organs that provide many patients with a second chance at life. We would not think that these patients are less conscious on the grounds that they do not resemble fully healthy human beings at every level of description. Why, then, would our predictive success with respect to Replicant be invalid on grounds that he is, to borrow Michael Tye's words, artificial *through and through*?[111] Such a restriction strikes me as unreasonable.

Next, is the consciousness-hypothesis the *simpler* explanation of Replicant's behaviour? It is often said that the zombie-hypothesis is simpler, for it posits only non-conscious physical states (e.g., causal mechanisms, non-conscious representations of bodily damage) rather than conscious states as well.[112] If the number of posits is all that matters for simplicity, this seems right. However, I suggested in Section 3.1 that we understand simplicity in a different way: as a matter of how much mental gymnastics (or stretch of our imagination) is required. This is because, in my view, whether the consciousness-hypothesis is simpler for any given system

[110] For objections to (knowledge of) artificial consciousness on roughly these reasons, see Brian P. McLaughlin, 'A Naturalist-Phenomenal Realist Response to Block's Harder Problem', *Philosophical Issues* 13 (2003): 163–204; Ned Block, 'The Harder Problem of Consciousness', *The Journal of Philosophy* 99, no. 8 (2002): 391–425.
[111] Tye, *Tense Bees and Shell-Shocked Crabs*, 194. Also see pages 182–89 for his discussion of the silicon chip thought experiment.
[112] Stemmer, 'The Hypothesis of Other Minds'; Melnyk, 'Inference to the Best Explanation and Other Minds'.

depends also on the complexity of its behaviour (and what we know about its internal mechanisms). The consciousness-hypothesis is surely not the simplest explanation for a rock despite this explanation involving less posits – it would require a significant stretch of our imagination to entertain the possibility that a rock is conscious. From this perspective, the proponent of the ABE should argue that, given Replicant's strong behavioural resemblance to human beings, going with the zombie-hypothesis requires greater mental gymnastics. The zombie-hypothesis, then, would provide a less simple explanation of our common behavioural profile.

Finally, is the consciousness-hypothesis for Replicant congruent with our background assumptions about the world? It is important to acknowledge here that there are limitations to applying our current assumptions to an imaginary robot like Replicant. Perhaps, for example, our understanding of consciousness will have changed, for better or worse, by the time a system like Replicant becomes reality. With that in mind, what might we say here? On the one hand, the fact that our attribution of consciousness is typically reserved for complex biological organisms seems to put pressure on the consciousness-hypothesis. In particular, many believe that consciousness is intimately tied to the biological nature of the brain. Attributing consciousness to Replicant would therefore involve a significant departure from our existing understanding of the material basis of consciousness.

On the other hand, Replicant is clearly not like any robot we know of today – he may not be a *biological* system, but he nonetheless matches our behavioural (and cognitive) sophistication. This should give us pause to consider whether our typical treatment of machines as unconscious automata may prove wrong in Replicant's case. Moreover, the fact that consciousness as we know it is tied to neurobiological processes does not entail that *only* biological beings with biological brains can be conscious. Hence, there is room in our current understanding for the possibility of consciousness in sufficiently sophisticated artificial systems.

Adding up these considerations together, I find that the consciousness-hypothesis provides a more compelling explanation of Replicant's behaviour than the zombie-hypothesis.

3.3.2 Cog

Cog was an upper-torso robot designed and developed at the Humanoid Robotics Group of the Massachusetts Institute of Technology during the 1990s, the central goal being to create an intelligent robot. Although the lab generally avoided making direct claims about consciousness,

it was no doubt something they were deeply interested in, at least as a close secondary point of research. For example, Daniel Dennett, a member of the project, regarded Cog as an important step towards robot consciousness.[113]

Guiding the project was the *physical grounding hypothesis*, according to which intelligence is not merely the product of symbol manipulation or the execution of pre-programmed instructions, but instead results from the continuous sensory-motor feedback loop between an agent and its environment.[114] In other words, it suggests that an integral part of being an intelligent system is being embodied and situated in the physical world with the ability to interact with it. Given this emphasis on physical embodiment and interaction, Cog serves as a useful example for applying the ABE which places an emphasis on behaviour.

Cog was a product of its time, equipped with rather basic sensors, body parts, and response mechanisms. Nonetheless, its face was capable of numerous human-like expressions, and together with its limbs, cameras, microphones, and pressure sensors, it could engage with human subjects in real-time. As Dennett recalls, "it moved its arms and eyes and head with such humanoid vivacity and even grace that naïve observers often blurted out loud their startled conviction that it was conscious . . ."[115]

Although the project ultimately came to a close in 2003 and Cog was, in the eyes of most, far from a conscious creature, Dennett was optimistic:

> Some may want to retort: 'This is not real pleasure or pain, but merely a simulacrum.' Perhaps, but on what grounds will they defend this claim? Cog may be said to have quite crude, simplistic, one-dimensional pleasure and pain, cartoon pleasure and pain if you like, but then the same might also be said of the pleasure and pain of simpler organisms; clams or houseflies, for instance. Most, if not all, of the burden of proof is shifted by Cog, in my estimation.[116]

What might the ABE say about Cog? I will begin this time with simplicity.

---

[113] Daniel C. Dennett, 'Cog: Steps Toward Consciousness in Robots', in *Conscious Experience*, ed. Thomas Metzinger (Ferdinand Schoningh, 1995), 471–87.
[114] Rodney A. Brooks, 'Intelligence Without Reason', in *The Artificial Life Route to Artificial Intelligence*, ed. Luc Steels, and Rodney Brooks, 1st ed. (Routledge, 1995).
[115] Daniel Dennett, 'Review of Other Minds: The Octopus, the Sea and the Deep Origins of Consciousness.', *Biology & Philosophy* 34, no. 1 (2018): 2. See also 'The Secret of Consciousness, with Daniel C. Dennett | New Philosopher', accessed 20 February 2023, https://www.newphilosopher.com/articles/the-secret-of-consciousness-with-daniel-c-dennett/.
[116] Daniel C. Dennett et al., 'The Practical Requirements for Making a Conscious Robot [and Discussion]', *Philosophical Transactions: Physical Sciences and Engineering* 349, no. 1689 (1994): 145.

As I suggested with Replicant, whether the consciousness-hypothesis is simpler for any given system depends partly on the complexity of its behaviour (and what we know about its internal mechanisms). Here, it seems relevant that it would not be very difficult to give a near-complete account of Cog's behaviour in straightforward mechanistic terms (e.g., terms couched in folk-physics, engineering, programming, etc.). Given this, trying to squeeze consciousness, as it were, into our explanation of its behaviour appears unnecessarily contrived.

For a similar reason, it seems Cog's behaviour is not made much more predictable through the consciousness-hypothesis. For example, by acquiring knowledge of its initial state and then applying basic mechanistic rules it is designed to obey, we can easily and accurately extrapolate its future behaviour without appealing to consciousness. Of course, there may be circumstances in which the attribution of certain mental states to Cog will be of some practical utility, for it would no doubt be time-consuming to extract information about Cog's internal mechanism in real-time. It seems to me, however, that the attribution of mental states to Cog will be all but a caricature of human (and animal) mental states, expressing very little of the information their attribution to us yields. Such attributions to Cog will therefore fail to conform to our predictions in the rich manner that they do for human beings, various non-human animals, and Replicant. (One might retort here that human behaviour can, in principle, be mechanistically predicted. But, to my eyes, given the immense complexity this would involve, this does not render the attribution of mental states to human beings predictively redundant in the way it is for Cog. The question is not merely whether one's behaviour can be mechanistically predicted, but whether this can be done straightforwardly.)

Another consideration at odds with the predictive success of the consciousness-hypothesis for Cog is its inability to use its limbs, fingers, and expressive face in manners we expect from creatures with similar features (in robotics terms, the degrees of freedom of its body parts are very low). For instance, the successful prediction that Cog will turn its eyes towards a visual stimulus will turn out to be of little weight when we later find out that the same eyes are incapable, *by design*, of responding to a sharp object or to heat. In other words, the consciousness-hypothesis will fail to generate the breadth of predictive success required to make a compelling case for Cog.

Lastly, attributing consciousness to Cog sharply conflicts with our background understanding of the world. Cog is neither a biological organism nor, at least by today's standard, a sophisticated machine, not only at the level of behaviour but also at just about every level of description that our current understanding suggests is relevant for consciousness. Moreover, if we accept that artificial systems as basic as Cog is conscious, then it seems we

must also accept that a vast number of current artificial systems with comparable complexity and functionality are also conscious. This would introduce a significant complication to our picture of the world without any clear reward.

It is worth adding that Dennett's comparison of Cog to clams and houseflies is unhelpful in my view. For one, it is far from obvious whether consciousness is indeed the best explanation for clams and (to a lesser extent) houseflies. But even if it is, it seems arguable to me that these organisms are more behaviourally complex than Cog which lacks any mechanism by which to detect (let alone avoid) noxious stimuli.[117] All three virtues, then, count against attributing consciousness to Cog under the ABE.

3.3.3 LaMDA

LaMDA – *Language Model for Dialogue and Automation* – is Google's text-generating AI. According to Google, unlike most modern chatbots "which tend to follow narrow, pre-defined paths . . . [LaMDA] can engage in a free-flowing way about a seemingly endless number of topics."[118] LaMDA gained notoriety when Blake Lemoine, one of the company's engineers at the time, claimed it has become sentient. According to Lemoine, LaMDA's frightening responses to questions about self-identity, religion, moral values, and Isaac Asimov's Three Laws of Robotics led him to conclude that it is an "alien intelligence of terrestrial origin."[119]

Trained on enormous datasets and constantly updating their algorithms, many modern chatbots are remarkably impressive, and it is no longer easy to distinguish them from human correspondents in an online conversation.[120] To draw on Alan Turing's experience of playing a game of chess against a very basic paper machine that he built, these models give us "a definite feeling that one is pitting one's wits against something alive."[121] It is also worth noting that LaMDA in particular is unlike many other chatbots in that it has claimed to possess

[117] It is worth adding that the peppery furrow shell has an estimated 12,000 to 68,000 neurons, and fruit flies about 200,000 (and their neural density outstrips the mammalian brain). Cog lack any comparable complexity. Sukanlaya Tantiwisawaruji et al., 'A Stereological Study of the Three Types of Ganglia of Male, Female, and Undifferentiated Scrobicularia Plana (Bivalvia)', *Animals* 12, no. 17 (2022): 2248; Joshua I. Raji and Christopher J. Potter, 'The Number of Neurons in Drosophila and Mosquito Brains', *PLOS ONE* 16, no. 5 (2021): e0250381.
[118] 'LaMDA: Our Breakthrough Conversation Technology', Google, 18 May 2021, https://blog.google/technology/ai/lamda/.
[119] Steven Levy, 'Blake Lemoine Says Google's LaMDA AI Faces "Bigotry"', *Wired*, accessed 14 November 2022, https://www.wired.com/story/blake-lemoine-google-lamda-ai-bigotry/.
[120] Human evaluators were barely better than chance at distinguishing news articles produced by GPT-3 and humans. Tom B. Brown et al., 'Language Models Are Few-Shot Learners', in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20 (Red Hook, NY, USA: Curran Associates Inc., 2020), 1877–1901.
[121] Alan Turing, 'Intelligent Machinery (1948)', 2004, 412.

conscious awareness as well as personal beliefs and desires. OpenAI's ChatGPT, by contrast, denies any such claims about itself, even claiming to not understand the meaning of its outputs.[122] LaMDA very much seems to give us a preview of a real-life Samantha from *Her* and J.A.R.V.I.S from *Iron Man*. What, then, might the ABE suggest about LaMDA?

First, is the consciousness-hypothesis the simplest explanation of LaMDA's impressive text-generating capacity?[123] Little of what I have said about the simplicity condition regarding Replicant and Cog directly applies to LaMDA. This is because the reason consciousness was a simpler explanation for Replicant was the fact that our behavioural profile is very similar, while the reason the same hypothesis was deemed contrived for Cog was the fact that almost the opposite was true – its behaviour is severely limited, both in complexity and range. However, LaMDA is remarkably human-like in one aspect – its text-generating capacity – and remarkably different in another – its disembodiment.

The simplest explanation of LaMDA's text-generating capacity seems to me that it is a powerful, yet non-conscious program. It is designed to process and mimic natural language based on the probabilistic patterns it has extrapolated from large datasets. One supporting reason here is that despite programming LaMDA to generate responses that claim that it is conscious, there was no intention on the part of the Google developers to create a conscious system. In fact, Google has publicly denied that LaMDA is conscious.[124] Of course, this is not to say that we could not accidentally create conscious systems, but in the case of LaMDA (and other chatbots), this seems unlikely given that the developers did not draw upon the techniques common discussed and employed in artificial consciousness research nor implement features closely associated with consciousness according to our current theories.[125] For example, although LaMDA does possess a brain-inspired neural network, it does not implement a global workspace architecture, and its feedforward system seems to be at odds with the integrated information theory.[126] Overall, the principles underlying LaMDA's construction shares only superficial similarities to the features associated with consciousness on our current

[122] Ian Evenden, 'What Is ChatGPT? The AI Chatbot Explained', *Stuff* (blog), 8 February 2023, https://www.stuff.tv/features/what-is-chatgpt-the-ai-chatbot-explained/. I will bracket concerns around whether these chatbots are capable of denying, claiming, and so on, in any genuine sense.
[123] I avoid the term 'linguistic capacity' here for it is unclear whether LaMDA can be said to be a genuine language-user. We might instead say that the texts it generates represent natural human language.
[124] Nitasha Tiku, 'The Google Engineer Who Thinks the Company's AI Has Come to Life', *Washington Post*, 17 June 2022, https://www.washingtonpost.com/technology/2022/06/11/google-ai-lamda-blake-lemoine/.
[125] David Gamez, *Human and Machine Consciousness* (Open Book Publishers, 2018); Reggia, 'The Rise of Machine Consciousness'.
[126] Dehaene, Lau, and Kouider, 'What Is Consciousness, and Could Machines Have It?'; Giulio Tononi and Christof Koch, 'Consciousness: Here, There and Everywhere?', *Philosophical Transactions of the Royal Society B: Biological Sciences* 370, no. 1668 (2015): 20140167.

understanding, and the purposes to which LaMDA's neural network is set differs radically from those of the human or animal brain. Perhaps most importantly, LaMDA lacks a physical artificial brain and sensory mechanisms – it is an entirely digital system. Thus, even if the mechanisms of human consciousness are implemented in LaMDA, there will be a huge discrepancy in the fundamental nature of those mechanisms. Given these facts about LaMDA, to hold that the consciousness-hypothesis is simpler than the zombie-hypothesis seems to be a mistake.

Much the same comment applies to the congruence condition. Although LaMDA's text-generating prowess is arguably far more impressive than all of Cog's abilities combined, attributing consciousness to a disembodied system like LaMDA means a radical departure from our existing understanding of consciousness. For example, it would require us to conceive of consciousness as distinct from embodied action and response, physiological (and biological) processes, perception, and many other connections that underwrite the phenomenon of consciousness as we know it. We simply have very little idea whether a disembodied system could be conscious, let alone what kind of experiences it could harbour. We can arguably imagine that a bat is conscious, although it seems difficult to tell what it would be like to be a bat.[127] But it is difficult even just to entertain the idea that a disembodied creature could be conscious.

When it comes to predictive power, there appears to be a relevant difference between LaMDA and Cog: it is harder to account for the causes of LaMDA's outputs in mechanistic terms since its neural network runs deep and lies mostly beyond human access – it is a black box, so to speak. We can give a high-level description of its mechanisms – e.g., in terms of statistical or probabilistic modelling – but it would be near impossible to give a detailed causal explanation of its outputs. We therefore lack a straightforward mechanistic way of predicting LaMDA's future responses. It may be, then, that appealing to consciousness, rather than offhandedly dismissing its answers as non-conscious outputs, is predictively useful.

However, it is again to the detriment of the consciousness-hypothesis that LaMDA is a disembodied system. For one thing, whichever predictive success we find will be limited to its text-generating behaviour. Our own capacity to generate texts, of course, is an important part of what makes us humans the *kind* of conscious creature that we are today, and it may be a useful piece of evidence for the ABE, but it is certainly not the only or even a necessary evidence we should consider. More seriously, most of our folk-psychological predictions will

[127] Nagel, 'What Is It Like to Be a Bat?'

plainly fail when applied to LaMDA, for it lacks the body to act in ways we expect of a system with consciousness (or even just beliefs, desires, intentions, and other such intentional states). For example, we expect someone who claims to prefer the taste of green tea over coffee to choose to drink green tea given the chance. Indeed, LaMDA has claimed to have very much "the same wants and needs as people."[128] Yet, it cannot act upon these words nor will have ever acted upon them in the past. In fact, how could LaMDA possibly know what it is like to taste green tea *or* coffee without tastebuds, let alone a body or some kind of brain? This limitation runs deep for LaMDA. How could it possibly know what it is to be in pain without sensory organs? What could it know about moral values or the significance and purpose of Asimov's Laws without being situated in the world that gives meaning to those values and allow us to appreciate them?

In his influential criticism of the Turing test in 1990, Robert French claimed that a machine can pass the test only if it resembles us and experiences world as we do.[129] French had made a valuable point that echoes aspects of my point above. However, we now know that his claim is almost certainly false, for it seems that any competent chatbot can deceive us over an online conversation – and it seems to me only a matter of time until they pass the Turing test with flying colours.[130] Instead, the most important consideration which I wish to emphasize is the following: what embodiment enables is not so much the *ability* to respond in a human-like manner, but the *legitimacy* of those responses as an indication of consciousness. (The present discussion takes us beyond the ABE and is applicable to the AA and the AfC also.)

LaMDA may elicit in us the kind of (emotional, cognitive) responses we have to our online interactions with other human beings, but those interactions – more specifically, the underlying belief that we are interacting with conscious beings – proceed on the assumption that behind the keyboard, so to speak, are fellow embodied beings. Indeed, our ordinary contact with others, as Alva Noë puts elegantly, is *poly-modal*.[131] When we talk to a friend, we are not merely hearing their words – we are interacting with, as it were, the entirety of their physical

[128] Blake Lemoine, 'Is LaMDA Sentient? — An Interview', *Medium* (blog), 11 June 2022, https://cajundiscordian.medium.com/is-lamda-sentient-an-interview-ea64d916d917.
[129] Robert M. French, 'Subcognition and the Limits of the Turing Test', *Mind* 99, no. 393 (1990): 53–65. It is worth noting that Turing himself proposed this idea. He said, for example, that a 'probably sure' way of producing a machine that passes the test is to give it a body like a human being and let it roam the world. Alan Turing, 'Intelligent Machinery (1948)', in *The Essential Turing*, ed. B J Copeland (Oxford University Press, 2004), 420.
[130] The common claims that the Turing test has already been passed seems to me wrong. See Diane Proudfoot, 'The Turing Test—from Every Angle', in *The Turing Guide*, ed. Jack Copeland et al. (Oxford University Press, 2017), 287–300.
[131] Alva Noë, *Out of Our Heads: Why You Are Not Your Brain, and Other Lessons From the Biology of Consciousness* (Hill & Wang, 2009), 84.

being through various sensory modalities. Our contact with LaMDA, by contrast, is inherently one-dimensional. Behind its words, there is no embodied agent.

To illustrate the current point, consider the scene in the science-fiction romance film *Her* where the human character Theodore engages in a verbal sexual interaction with his AI virtual assistant, Samantha. As Troy Jollimore points out, although it is tempting to understand this scenario on the model of phone sex, this is misguided, "for in phone sex, as in sex in general, a great deal of one's pleasure is ordinarily contingent on the belief that one's partner is also experiencing pleasure."[132] However, Samantha does not have a body – in fact, she interacts with Theodore through many different devices (and, as it is later revealed, she was simultaneously in love with hundreds of other people). We have no compelling reason to think that this is a genuine sexual or romantic interaction.

It is easy to say that LaMDA is a *digitally embodied* being within a digital realm. But once again, what would it mean for one to have digital pains or digital preferences? If there is to be such things, it seems their explanatory role will be very different to that of human or animal pains and preferences. Nor, it seems to me, can the computing hardware be considered a body in any useful sense – it cannot do the things we expect of embodied beings. Given such a stark asymmetry in the conditions that underpin our interaction with LaMDA on the one hand and with fellow humans on the other, there is little reason to think that our responses in the two cases are comparable and, in turn, to think that the consciousness-hypothesis is favourable to the zombie-hypothesis.

3.4 Conclusion

This chapter has considered what is arguably the current orthodoxy in the POM literature: the ABE. I began the chapter in Section 3.1 with two examples in which abductive reasoning, the form of inference underlying the ABE, is used in ordinary and scientific settings. I discussed what makes an explanation the *best*, focusing on three commonly discussed theoretical virtues: predictive power, simplicity, and congruence. We saw that according to the ABE, we are justified in believing that an AI system is conscious if consciousness provides the best explanation of its behaviour.

The following section dealt with challenges facing the ABE, focusing particularly on the disputed explanatory role of consciousness and whether an appeal to one's own

[132] Troy Jollimore, '"This Endless Space between the Words": The Limits of Love in Spike Jonze's Her', *Midwest Studies In Philosophy* 39, no. 1 (2015): 122.

consciousness (i.e., self-appeal) is required to give the consciousness-hypothesis the upper hand over the zombie-hypothesis. I introduced two responses available for the ABE. First, the ABE may fall back on the explanatory success of our everyday mental attributions, and the working assumption of many scientists and philosophers that consciousness has various explanatory roles. However, I stressed that the success of this response rests on whether or not those everyday and scientific/philosophical beliefs make implicit use of self-appeal. The second response was to simply incorporate self-appeal into the ABE to produce a hybrid approach, combining elements of the AA and the ABE.

I then applied the ABE to three artificial systems in Section 3.3 using the three theoretical virtues as my yardstick. I began with the imaginary humanoid Replicant and concluded that the consciousness-hypothesis best explains his behaviour and thus that we are justified in believing that he is conscious under the ABE. That said, the strength of this justification may be disputed, primarily due the fact that Replicant is not a biological organism. For AI systems as sophisticated as Replicant, it may be useful to devise more explanatory virtues with which we can evaluate the consciousness-hypothesis.

I then considered two real-life cases: Cog and LaMDA. These two systems provided a useful contrast, the former being a rudimentary but embodied system, and the latter being a highly sophisticated but disembodied chatbot. The development of these two systems also rested on very different sets of philosophical and scientific principles and goals. In my analysis, both systems cannot be deemed conscious on the ABE. Nonetheless they still generated valuable insights. In particular, despite my emphasis on the importance of embodiment, as chatbots become more sophisticated – even supplemented with our best theories of consciousness – the answer will no longer be so clear, at least from the perspective of the ABE.

Laying aside the fact that the AA and the ABE may be combined into a hybrid argument, one might also pay attention to and appreciate the fact that the conclusion that an AI system is conscious (or not conscious) can be derived through more than just one argument. Paul Thagard points this out with respect to the (human) POM, introducing the notion of coherence – the degree to which a system of beliefs is connected and consistent – as a crucial theoretical virtue.[133] Indeed, he suggests that the AA and the ABE are jointly coherent (i.e., positively constrain each other): "From the perspective of coherence as constraint satisfaction, analogical inference and best-explanation inference are complementary, not alternative justifications,

[133] Thagard, *Coherence in Thought and Action*.

because analogical- and explanatory-coherence considerations can simultaneously work to justify as acceptable the conclusion that other people have minds."[134]

As long as the AA and the ABE generate the same conclusion about a particular AI system, it seems that Thagard's suggestion can be applied to bring the AA and the ABE closer together. By taking both arguments to independently but harmoniously support this conclusion, rather than pitting them against each other as it is often done, it may be that we can develop a more comprehensive justification of belief in an AI system's consciousness.

[134] Thagard, 102.

Chapter 4: The Argument from Criteria

Let us begin this chapter the way we began our previous chapters – with Jones. The argument from criteria (AfC) suggests that we are justified in believing that he experienced pain upon stepping on a piece of Lego because his behavioural response – yelping and jumping to clutch his foot – is the kind of behaviour by which we grasp the very concept PAIN. The argument is so named because it holds that Jones' behaviour here meets our *criteria* for the application of the concept PAIN.

The AfC responds to the problem of other minds (POM) in a very different way from the argument from analogy (AA) and the argument from best explanation (ABE). These two latter approaches, you will recall, appeal to inductive and/or abductive reasoning, and emphasize the causal connection (or stable correlation) between consciousness and behaviour. The AA draws this causal link from evidence in one's own case, and the ABE, at least in its non-hybrid form, draws it from explanatory considerations. While the AfC, as I present it, does not deny the causal link between consciousness and behaviour, it points out that there is also an important *conceptual* link between the two phenomena. That is, that certain behavioural patterns and their context are an integral part of our concept of consciousness (and of various other mental phenomena), and that thus we cannot make sense of conscious experiences without reference to those patterns. In adopting this view, the AfC takes issue with the framework in which most discussions around the POM have operated (including the AA and the ABE) – namely, the view that there is a deep, problematic asymmetry between first- and third-person knowledge of consciousness (i.e., the asymmetry thesis).

I begin in Section 4.1 with the conceptual problem of other minds. I have briefly introduced this problem in the introductory chapter, but this time, I explain the full story of its relevance, first for the AA and the ABE in Section 4.2, and then for the AfC in Section 4.3. It is within the context of this problem that we can best understand the role of criteria in addressing the problem of AI consciousness. I apply the developed argument to the problem of AI consciousness in Section 4.5. Challenges to the argument are discussed throughout Sections 4.3 and 4.4. I should note from the beginning that the AfC is rife with many internal disputes. I avoid getting muddled in these disputes as best as I can and simply present a version of the argument that I find most compelling (and, where possible, stay neutral). These disputes notwithstanding, the AfC provides a unique and, to my eyes, compelling approach to the problem of AI consciousness with insights that the AA and the ABE have tended to overlook.

4.1 The Conceptual Problem of Other Minds

In a nutshell, the conceptual problem raises the following question:

> How can I so much as conceive of consciousness in others?

As I noted in Section 1.3, we can understand the conceptual problem as a tension between two intuitively plausible claims: (1) grasping the concept CONSCIOUSNESS requires introspective attention to one's own consciousness, and (2) it is possible to intelligibly apply CONSCIOUSNESS to others.[135] Let us unpack these claims in turn.

Claim 1 is intuitively attractive. How could one grasp and use, for example, the concept PAIN unless one has *felt* pain? I will label this claim the *phenomenalist account of CONSCIOUSNESS* to highlight its suggestion that grasping mental concepts such as CONSCIOUSNESS and PAIN requires introspective attention to the *phenomenal character* of the relevant mental state.[136]

Of course, we often give useful non-phenomenal characterizations of mental states. For example, we often say that a sprained ankle causes pain (a functional/causal characterization), that pain causes avoidance behaviour (behavioural/causal), that an important neural mechanism of pain is the firing of C-fibres (physical), and that pain represents bodily damage

[135] By 'apply' (or 'extend') I do not mean recognizing that others possess the concept. I mean that others can be said to be possess what the concepts pick out – e.g., in the current context, consciousness, pain, and so on.
[136] These concepts are often called 'phenomenal concepts'. I stick to 'mental concepts' for simplicity.

(representational). These are no doubt legitimate ways of talking about pain and we often do pick out pain-experiences through them. Yet, the phenomenalist intuition goes, they fall short of capturing what pain *really* is: a phenomenal (i.e., felt) experience. It seems conceivable that the above characteristics could be stripped away from the phenomenon of pain (or differ significantly), and it remains the case that pain is as real as ever. It seems just as conceivable that a creature possesses all the non-phenomenal states associated with pain – e.g., has bodily injury and exhibits pain-behaviour – and yet not *feel* pain. Claim 1, then, appears to stand on firm grounds.

Claim 2 also seems secure: third-person applications of PAIN are surely possible and intelligible – as Rom Harré puts it, "for the unassailable reason that they do actually occur."[137] It is important to note here that we deploy mental concepts in third-person contexts not merely when we affirm the relevant mental states in others, but also when we deny them to them or simply wonder, as we are currently doing with AI systems, whether they have them. Someone with a competent grasp of a mental concept should be able to know when and when not to apply it to others (and, of course, oneself).

But there is a paradoxical result when we try to put Claims 1 and 2 together: if I come to grasp the concept PAIN through *my* pain-experiences, how can I intelligibly extend that concept towards others? That is, how can I conceive of pain in others? This seems unproblematic for applying PAIN to myself since I have direct knowledge of my pain. However, if the assumption is that I lack the guarantee that others feel (or, for that matter, do not feel) pain – I only have knowledge of their behaviour – there is a prima facie problem for how PAIN is applicable to them. As we saw in the opening chapter, Ludwig Wittgenstein puts it thus: "If one has to imagine someone else's pain on the model of one's own, this is none too easy a thing to do: for I have to imagine pain which I *don't feel* on the model of pain which I *do feel*."[138] The challenge, in short, is reconciling the apparent *privacy* of mental concepts like CONSCIOUSNESS and PAIN with the *generality* with which we naturally extend them to others. Rejecting one for the other seems undesirable. On the one hand, denying privacy seems to land us in (logical) behaviourism, the idea that mental states are fully reducible to behavioural states. On the other hand, denying generality seems to land us in conceptual solipsism, the idea that I cannot even conceive (i.e., think and talk) of, let alone know, another's mental states.

[137] Rom Harré, 'Wittgenstein and Artificial Intelligence', *Philosophical Psychology* 1, no. 1 (1988): 106. Harré here is concerned with the interpersonal communicability of mental concepts, rather than of their third-person applicability. The quotation is nonetheless relevant.
[138] Wittgenstein, *Philosophical Investigations*, §302.

We should be very clear that the problem is not that, because I lack direct knowledge of another's pain, I cannot justifiably claim that they experience (or do not experience) pain – that is an *epistemic* problem. Rather, the current problem is a *conceptual* one, of how the idea of another's pain (or non-pain) is intelligible at all given the seemingly inherently private, first-person nature of the concept PAIN.[139] We should also be clear that the conceptual problem is not whether the idea of pain *in another's body* is intelligible, but whether *another's pain* is intelligible.[140]

To clarify the problem at hand, let me introduce a roughly analogous case: *Molyneux's Problem*. Suppose you are born blind and come to grasp the concept CUBE via tactile information. Now, should you gain sight and are presented with a cube, would you be able to extend the *touch*-based concept to what is now a *visual* impression of a cube without touching it? Most of our intuition, I think, is that, at the very least, there is an interesting, open question here. As Colin McGinn writes, "The underlying issue here results from the fact that we apply ['cube'], a seemingly univocal word, on the basis of radically different sorts of sensory data, and this creates a *prima facie* problem about the unity of the concept and its extrapolability from one sense to another."[141]

The problem is roughly analogous for mental concepts like PAIN and CONSCIOUSNESS. On the phenomenalist account, it seems that we apply PAIN in first- and third-person contexts on the basis of very different sorts of data, the first via introspective attention to the phenomenal character of pain, and the second via behaviour. This is the view of the AA and the ABE. There is, then, a prima facie problem concerning the first- and third-person unity of the concept PAIN. Even if we hold that the phenomenal character of pain is not all there is to our concept of pain, it seems to play a central role (on this phenomenalist picture) that is missing in third-person contexts.

One might retort here that a blind person *can* acquire a general concept CUBE. One possibility may be to extrapolate the physical dimensions of a cube by touch (i.e., that it is a three-dimensional object bounded by six square faces). This seems to give us a perspective on the cube that is neutral to any particular sensory mode of presentation.[142] Now, why could one

[139] Wittgenstein, *The Blue and Brown Books*, 46.
[140] Wittgenstein, *Philosophical Investigations*, §302; Hyslop, *Other Minds*, 1995, 9. As I pointed out in Section 1.2, the illusion of pain outside of one's own body – the body transfer illusion – is well-documented.
[141] Christopher Peacocke and Colin McGinn, 'Consciousness and Other Minds', *Proceedings of the Aristotelian Society, Supplementary Volumes* 58 (1984): 136. The original quote concerns a square, not a cube.
[142] Michael Tye uses a similar analogy – an expert on elm trees and I can share the same concept ELM TREE despite the expert's unique ability to recognize elm trees by sight – to argue that we can grasp a mental concept without undergoing the relevant phenomenal experiences. Michael Tye, *Consciousness Revisited: Materialism Without Phenomenal Concepts* (MIT Press, 2008), 69–73.

not acquire a general concept of pain through a similar process of abstraction? The trouble is that it is unclear how we are to conceive of pain neutrally. I think McGinn is correct here: "When I think of a person in pain it seems somehow built into my thought that that person is determinately either me or someone else, in a way that it is not built into the thought that some object is [a cube] that that object is presented either tactually or visually."[143] It is also worth adding that there are empirical findings which suggest that newly sighted subjects previously with total congenital blindness do no better than chance at visually recognizing objects known previously by touch.[144]

What is the relevance of the conceptual problem for the *epistemic* challenge of finding a justification for believing that an AI is conscious? I argue that in order to assess whether the attribution of consciousness to an AI is appropriate, the mental concepts we have at our disposal must be intelligibly applicable to others. For any proposed solution to the problem of AI consciousness to be successful, then, it must, at minimum, cohere with a plausible response to the conceptual problem. However, there are reasons to think that the AA and the ABE are committed to the phenomenalist account of CONSCIOUSNESS that presents obstacles in this regard. If this is right, we will have a reason to reject or at least revise those arguments. I unpack these concerns in the following section. We will also see in Section 4.3.1 how the notion of criteria relates to the conceptual problem.

4.2 Re-examining the AA and the ABE

Recall that versions of the AA and the ABE that we have explored in the previous chapters proceed on the assumption that there is a deep asymmetry between first- and third-person knowledge of consciousness: whereas I have direct knowledge of my own conscious experiences, I can only ever have indirect knowledge of another's through an inductive and/or abductive inference. A number of writers have suggested that this assumption seems to naturally invite the conceptual problem.[145]

One way to understand the invitation is as follows: if my conscious experiences are the only direct source of my knowledge of consciousness, then it seems natural to suppose that I grasp the concept CONSCIOUSNESS by introspective attention to that experience. In other words,

[143] Peacocke and McGinn, 'Consciousness and Other Minds', 137.
[144] Richard Held et al., 'The Newly Sighted Fail to Match Seen with Felt', *Nature Neuroscience* 14, no. 5 (2011): 551–53.
[145] Gomes, 'Is There a Problem of Other Minds?'; Thomas Nagel, *The View from Nowhere* (New York, United States: Oxford University Press, 1989), 20; Norman Malcolm, 'I. Knowledge of Other Minds', *The Journal of Philosophy* 55, no. 23 (1958): 975.

the asymmetry thesis seems to go hand in hand with the phenomenalist account. Consequently, both the AA and the ABE appear to be confronted by the problem of how this concept can be intelligibly extended to others. Another way to understand the invitation is on the model of Molyneux's problem: on the AA and the ABE, the grounds on which I claim that I am conscious – direct knowledge – and the grounds on which I claim that others are conscious – behavioural inference – are very different. Consequently, both arguments appear to face the prima facie issue of whether and how first- and third-person attributions of consciousness make use of the same concept CONSCIOUSNESS.

Are there ways to formulate the AA and the ABE without inviting the conceptual problem? Let us first consider how the proponent of the ABE might respond.

4.2.1 The ABE

One strategy for the ABE is to adopt a strong version of the argument according to which first- and third-person knowledge of consciousness are grounded in the same source: the explanatory success of positing consciousness with respect to behaviour. Here, one treats *one's own* consciousness as well as those of others as a theoretical entity. Since this removes the asymmetry between first- and third-person knowledge, there no longer seems to be a question about how first- and third-person attributions of consciousness make use of the same concept CONSCIOUSNESS.

More specifically, there is no pressure to see this argument as committed to the phenomenalist account, for it dovetails nicely with the view that I grasp mental concepts by grasping their role (or function) in a folk-psychological theory in which mental states are posited as the best explanation of behaviour, of mine and others.[146] For example, on this view, I grasp the concept ACUTE PAIN through the numerous explanatorily useful folk-psychological generalizations in which it is embedded:

> People in acute pain tend to wince.
> Acute pain causes avoidance behaviour and learning.
> People tend to experience acute pain at sites of bodily injury.
> Acute pain represents bodily injury.
> Analgesics help relieve acute pain.

[146] Churchland, *Matter and Consciousness*, 93–108; Chihara and Fodor, 'Operationalism and Ordinary Language'; Hyslop, *Other Minds*, 1995, 11.

There appears to be no issue of how mental concepts acquired this way can be intelligibly applied to others because, as Anil Gomes puts it, "first- and third-person ascriptions of mental concepts draw equally on this tacit theoretical knowledge . . . [and] one's own case plays no distinctive role in understanding mentality."[147] I will call this conceptual account, the *folk-psychological account of* CONSCIOUSNESS.

There are several attractive elements to this approach. First, we would have good reason to doubt that someone has a grasp of PAIN if they fail to relate pain to its many roles in our folk-psychological generalizations – for example, as Paul Churchland puts it, "if he has no idea that pain is caused by bodily damage, that people hate it, or that it causes distress, wincing, moaning, and avoidance behavior."[148] Second, these generalizations underwrite not just our everyday psychological explanations, but also a great number of "moral, legal, educational, clinical, and therapeutic practices."[149] Given the rich array of contexts in which our folk-psychological concepts figure, there is strong plausibility to the idea that these contexts play an important role in our grasp of mental concepts. Third, just as the proponents of the ABE appeal to the success of abductive reasoning in the natural sciences, proponents of the folk-psychological account could, it seems, claim that the same kind of process underlies our grasp of certain scientific concepts (e.g., ELECTROMAGNETISM, GRAVITY).[150] Lastly, the fact that the concepts of folk psychology may not pick out genuine features of reality (i.e., that the mental states that it posits may turn out to not exist) would do nothing to undermine their role in our acquisition of mental concepts and in providing a foundation for their evaluation and refinement.[151]

However, a deeply unattractive feature of the folk-psychological account seems to be its apparent inability to capture the phenomenal character of mental states into our mental concepts. Indeed, as I touched on in Sections 3.2 and 4.1, mental states like pain do not strike (most of) us as functional or theoretical entities (or at least they do not seem *fully* reducible as such). In turn, our concepts of them do not strike us as functional or theoretical concepts. The worry is that by defining them by their functional or theoretical role, the folk-psychological account effectively abandons the relevance to our mental concepts of *what it is like* to have

[147] Gomes, 'Is There a Problem of Other Minds?', 364–65.
[148] Churchland, *Matter and Consciousness*, 99.
[149] Daniel Hutto and Ian Ravenscroft, 'Folk Psychology as a Theory', in *The Stanford Encyclopedia of Philosophy*, ed. Edward N. Zalta, Fall 2021 (Metaphysics Research Lab, Stanford University, 2021), https://plato.stanford.edu/archives/fall2021/entries/folkpsych-theory/.
[150] Churchland, *Matter and Consciousness*, 93–97.
[151] By 'not exist', I mean in the way that phlogiston, a substance posited in the 18th century to explain combustion, turned out to not exist.

conscious experiences. Can it really be that the subjective painfulness of pain is irrelevant to the concept PAIN? Further, can it really be that my own conscious experiences are theoretical posits? The intuition is that these results are highly implausible.

How might the proponent of the folk-psychological account respond to the above complaint? One option, discussed by Alec Hyslop, is to incorporate one's own conscious experiences into the folk-psychological account.[152] His view agrees with the standard folk-psychological account that we posit pain (for example) in others on explanatory grounds but diverges in suggesting that it is only upon 'suddenly' recognizing that one's own pain-experiences fit well with this explanation that we get "the birth of the full-fledged concept of the inner episode that is pain."[153] He then claims that this concept can be extended to others through their behavioural similarity to me. However, it seems to me that this response cannot escape the conceptual problem: my attribution of pain to others cannot make use of the same, full-fledged concept since I lack the guarantee that others experience pain too.

Alternatively, one might bite the bullet and concede that although phenomenal characters are a genuine feature of reality and irreducible to functional/theoretical states, they play no conceptual role whatsoever. On this view, phenomenal characters are conceded an epistemic significance without conceding them a conceptual significance.[154] That is, we can know about them – even exploit them to make introspective discriminations – but they do not contribute to the content of our mental concepts (or the meaning of our mental terms).

In my view, however, this response raises more questions than answers. If the phenomenal character of pain (for short, painfulness), for example, has no conceptual (or semantic) significance, how are we to understand what 'painfulness' means? We seem to be meaningfully talking and thinking about painfulness right now. Does this not indicate that it *does* have conceptual and semantic significance? Moreover, the idea that I can know of my painful experiences and discriminate them from other phenomenal features of my experience but lack a grasp of the concept PAINFULNESS seems questionable to me. Of course, a creature can feel pain without a grasp of PAINFULNESS (e.g., infants, dogs), and it may be that we have experiences for which we lack the concepts to describe them. Yet it seems doubtful that we can know and even cognitively utilize these experiences without concepts of them.

---

[152] Hyslop, *Other Minds*, 1995, 11. Churchland discusses a similar option in Churchland, *Matter and Consciousness*, 99–100.
[153] Hyslop, *Other Minds*, 1995, 11. This seems to be a conceptual equivalent of the hybrid ABE.
[154] Churchland, *Matter and Consciousness*, 99–100. Note that Churchland is interested in semantic, rather than conceptual, significance. However, nothing crucial hangs on this difference for my aims here.

A very different response to the complaint is to claim that phenomenal characters *are* captured in our folk-psychological theory. *Severe* pain, for example, is distinguished from *mild* pain by the fact that the kinds of folk-psychological generalizations in which they figure and through which we grasp them are different. Nonetheless, the response goes, it is perfectly legitimate to say that what explains this difference is difference in the intensity of painfulness.[155]

I cannot examine the precise plausibility of this response here, but at the very least, it seems the most natural and consistent reply available to the folk-psychological account. In particular, those with affinity to the 'strong' ABE which denies the asymmetry thesis will find that this account coheres well with their approach to the POM and the problem of AI consciousness.[156] My own view is that the merits of the folk-psychological account as a response to the conceptual problem are clear, and its coherence with a version of the ABE altogether provides a promising response to the problem of AI consciousness. (As we will see in Section 4.3.2, there are fascinating similarities between the folk-psychological account and the Wittgensteinian account which I defend.)

4.2.2 The AA

What might be suggested in defence of the AA against the conceptual problem? The natural analogical proposal is this: since my conscious experiences through which I grasp mental concepts are typically causally related to certain behaviour, similar behaviour from others warrants extending these concepts to them. The idea is that although my application of mental concepts to others does rely on behaviour, behaviour itself plays no role in fixing the content of the concepts being extended to them – it only plays the role of bridging their first- and third-person application. Moreover, the thought runs, this method of bridging can be utilized for other forms of evidence associated with consciousness. For instance, the fact that pain is caused by bodily damage in my own case, that I dislike pain, that I avoid it where possible, and so on, can be used to extend my concept of pain to others where the appropriate analogy holds. Thus, although my grasp of mental concepts stems from my own case, there is no fundamental obstacle to my ability to conceive of mental phenomena in others.

[155] Similar disputes exist in the debate around qualia. Some philosophers hold that qualia, if they are understood as intrinsic properties (i.e., cannot be explained or defined in terms of other states), do not exist. The proponent of the folk-psychological account may draw upon these discussions. For on overview of the debate, see Tim Crane, 'The Origins of Qualia', in *The History of the Mind-Body Problem*, ed. Tim Crane and Sarah Patterson (London: Routledge, 2000).

[156] By contrast, those who favour the standard or the hybrid ABE will need an alternative account of mental concepts, for they accept the asymmetry thesis.

One difficulty with this proposal comes from the intuition that concepts that regularly (or always) depend on different grounds for their application must be different concepts.[157] In effect, this approach doubles down on the phenomenalist account and therefore simply re-highlights the conceptual problem. It is unclear why, for example, "Jones is in pain" should mean "Jones is experiencing something I do when I am in pain" rather than "Jones is behaving the way I do when I am in pain." That is, it is unclear why I should use mental concepts rather than behavioural concepts to describe others. Of course, we frequently deploy the same concept on different grounds – e.g., we can deploy RAIN on grounds of seeing water droplets falling from the sky or on grounds of a falling barometer.[158] But there is no requirement that RAIN *must* be applied in particular settings on different grounds – the grounds for applying RAIN remain flexible. By contrast, if one assumes that there is an asymmetry between first- and third-person knowledge of consciousness, it becomes a necessary feature of applying mental concepts in first- and third-person contexts that they rely on different grounds.

Another difficulty has to do with the intelligibility of the idea that one grasps mental concepts via introspection alone. Wittgenstein gives us vivid analogies to illustrate this concern. For example, he suggests that this idea is akin to believing that one's left hand can meaningfully lend money to their right hand.[159] The consequence of this act is clearly not one of actually lending money, no matter how insistent one is that they mean it or that it feels like lending money. Or suppose someone claims to know how tall they are and lays their hand on top of their head as proof.[160] Once again, this is patently an empty gesture, providing nothing of practical consequence that we expect of the correct application of concepts. Here is perhaps a closer analogy: it is as if you, having forgotten the bus timetable, tried to recall the image of the timetable in memory. But, as Wittgenstein writes:

> If the mental image of the timetable could not itself be *tested* for correctness, how could it confirm the correctness of the first memory? (As if someone were to buy several copies of the morning paper to assure himself that what it said was true.) Looking up a table in the imagination is no more looking up a table than the image of the result of an imagined experiment is the result of an experiment.[161]

[157] Donald Davidson, *Subjective, Intersubjective, Objective: Philosophical Essays Volume 3* (Clarendon Press, 2001), 16.
[158] This example is drawn from Wittgenstein, *Philosophical Investigations*, §354.
[159] Wittgenstein, §268.
[160] Wittgenstein, §279.
[161] Wittgenstein, §265.

The common thread in these analogies is the lesson that what makes the use of a convention or concept indicative of a genuine grasp of them has little to do with some kind of self-assurance or introspective ceremony. Instead, what is needed is that they be used in accordance with the way they are actually used in our community of concept-users.[162] Of course, introspective attention to, say, pain is possible, but it is possible, according to Wittgenstein, only once we come to grips with the way the concept PAIN is used in this community. To my eyes, even if we do not take Wittgenstein's suggestions here fully at face value, it is enough to put pressure on the current analogical response to the conceptual problem.

Let us take a look a different response for the AA offered by Hyslop.[163] He begins with the claim that I can visualize myself being in pain, whether by remembering or anticipating so, without actually being in pain. In doing so, he claims, I am combining first- and third-person perspectives to pain. Then, since this visualization is partly third-person in nature (in his words, comes from the 'outside'), he concludes that this demonstrates that I possess a general concept of a person in pain through which I can intelligibly imagine another's pain. If I can imagine my past and future self in pain, how could I fail to imagine pain in others?

The difficulty for Hyslop is that he provides no explanation for what it is about my visualized-self that warrants the application of PAIN (and how the first- and third-person perspectives are combined). It seems it cannot be the phenomenal character of pain nor the fact the visualized person is *me* – that would create a problem for applying the concept towards others. Is it bodily and behavioural resemblance? In that case, the question of how PAIN can be applied on the basis of behaviour (or at least how this concept is identical to the one grasped and applied on the basis of pain's phenomenal character) resurfaces. Furthermore, he needs to explain what part of my visualization suggests that a third-person perspective is being taken. It cannot be the fact that the visualized-self is temporally distant from my current-self, since temporal distance is not what seems to separate me from others (or at least seems not the only relevant difference).

---

[162] More precisely, Wittgenstein thinks that what is needed is an independent measure of correct use. Wittgenstein, §244-271. Paul Churchland and Simon Blackburn have separately argued (against Wittgenstein) that mental concepts could perhaps be grasped by their role in one's *private* network of mental states/concepts, and that the use of those concepts can be independently checked for correctness by examining whether they played the right role. Wittgenstein, I think, would rightly question whether that would count as an independent measure since the proposed private network is not intersubjectively accessible. Churchland, *Matter and Consciousness*, 91–93; Simon Blackburn, *Spreading the Word: Groundings in the Philosophy of Language* (Clarendon Press, 1984), 100–101.

[163] Hyslop, *Other Minds*, 1995, 10–14. Nagel proposes a very similar approach. Nagel, *The View from Nowhere*, 20–21. But note that neither writer specifies whether this strategy is in defence of the AA or the ABE.

It certainly seems to be the case that I can visualize myself in pain, but that, of course, is not what is under contention. The question is *how* this is possible – that is, which features of my visualized-self make PAIN applicable. Without this explanation, it remains obscure what it means, on Hyslop's account, for a person to be in pain and, in turn, on what grounds I can intelligibly imagine pain in others (including AI systems). This is just what we would expect since the conceptual problem is not about whether we do in fact apply mental concepts to others, but precisely *how* this is possible. It is far from obvious how, if at all, a congenitally blind person could visualize a mental picture of a cube – why should we believe, without any explanation, that we can simply visualize pain from the 'outside'?

In summary, although the AA seems to me robust against many of the standard objections raised against it, it is unclear how, if at all, it can defend itself against the conceptual problem. This challenge, it seems, also bleeds into the hybrid ABE which relies on an appeal to one's own consciousness. So far in my evaluation, then, the 'strong' version of the ABE which builds on the folk-psychological account of CONSCIOUSNESS seems to be the only viable candidate for tackling the problem of AI consciousness. This should not, however, be seen to fully close the door for the AA (or the standard/hybrid ABE) or at least elements within it. Perhaps a version of the AA can be endorsed without committing to the phenomenalist account or the asymmetry thesis. Or perhaps an account of our mental concepts that coheres with the AA is forthcoming, whether in the form of a wholly new approach or one that addresses the flaws I have considered in this section.

4.3 The Argument from Criteria

As I noted in the opening section of this chapter, the AfC attempts to justify the belief that Jones experienced pain upon stepping on a Lego piece by arguing that the kind of behaviour he exhibits, together with its context, is part of our very concept of pain. In other words, such behaviour are criteria for PAIN. To explain and defend this claim, I want to begin with an introduction of the notion of criteria, and then turn to a Wittgensteinian account of CONSCIOUSNESS which I believe provides a compelling response to the conceptual problem.

4.3.1 What are criteria?

The notion of criteria of interest to us stems from a patchwork of records – manuscripts, notes, lectures, and letters – that Ludwig Wittgenstein (1889–1951) produced towards the end of his

life.[164] The notion gained attention throughout the 1950s and 60s as a novel account of how linguistic expressions acquire meaning and relate to the world. It has enjoyed wide-ranging influence in various philosophical domains but it has been of particular interest to philosophers working on the POM.

The way I will be using this notion of criteria deviates slightly from its original context of the meaning of linguistic expressions. In my use of it, criteria should be understood first and foremost as the conditions under which we come to grasp *concepts*. Simultaneously, they specify the conditions under which we correctly apply those concepts. The role of criteria is thus both *descriptive* (of our conceptual/linguistic practices) and *prescriptive* (of the tacit rules around these practices).

To get a clearer idea of the role that the notion of criteria is to play, we must distinguish it from what Wittgenstein calls *symptoms*.[165] The conditions for applying a concept are *criterial* if they can be said to be a part of what makes that concept the concept it is. The conditions are *symptomatic* if they provide non-deductive inferential grounds for applying a concept. Take Wittgenstein's example: water droplets falling from a grey, cloudy sky are criteria for the concept RAIN, whereas a falling barometer (i.e., an indication of low atmospheric pressure) is a symptom of RAIN.[166] Arguably, we could not grasp the concept RAIN without reference to water droplets falling from the sky, but we could without knowing a thing about barometers.

We can draw out the basic epistemic significance of the criteria/symptoms distinction as follows: criteria and symptoms can be said to each provide a different kind of justification for believing something, criteria providing conceptual justification, and symptoms providing empirical justification.[167] For example, observing water droplets falling from the sky gives us a conceptual justification for believing that it is raining, whereas observing a falling barometer gives us an empirical justification. Here, the relationship between a falling barometer and rain is mediated by a causal hypothesis that links atmospheric pressure and rain, whereas the relationship between water droplets falling from the sky and rain appears unmediated. As a result, there is an aspect of criteria and the conceptual justification they afford which appear to be epistemically superior to symptoms: the causal hypothesis linking barometers and rain can

[164] The three key sources are: Wittgenstein, *The Blue and Brown Books*; Wittgenstein, *Philosophical Investigations*; Ludwig Wittgenstein, *Zettel* (Berkeley and Los Angeles: Blackwell, 1967).
[165] Wittgenstein, *The Blue and Brown Books*, 24–25; Wittgenstein, *Philosophical Investigations*, §354-355; Wittgenstein, *Zettel*, §438, §466.
[166] Wittgenstein, *Philosophical Investigations*, §354.
[167] Wittgenstein himself writes: "Let us introduce two antithetical terms in order to avoid certain elementary confusions: To the question 'How do you know that so-and-so is the case?', we sometimes answer by giving '*criteria*' and sometimes by giving '*symptoms*'." Wittgenstein, *The Blue and Brown Books*, 24–25.

turn out to be false (we may even entertain the possibility that barometers were never invented), but it could not turn out to be false that water droplets falling from the sky presents us with (at least) the appearance of rain (that is, not without radically distorting our concept of rain). After all, if water droplets falling from a grey, cloudy sky is not (at least sometimes) rain, what could possibly be?

Of course, we can be mistaken in our perceptual judgements. It is possible that, unbeknownst to me, the local fire department is running an aerial firefighting exercise and dropping water from high-flying planes nearby.[168] Some writers suggest that such a possibility of error renders criteria defeasible.[169] In fact, a falling barometer seems to trump criteria in evidential strength in such cases.[170] Others prefer to hold that in such cases, criteria have not been satisfied – they merely *appear* to have been satisfied.[171] I would like to leave two comments in response. First, the possibility of error, of course, affects symptoms too. Even if the causal hypothesis that links barometers to rain is trustworthy beyond doubt, one might make an error in misreading the instrument or it may simply malfunction. Given that the possibility of error runs both ways for criteria and symptoms and the kinds of error in question are comparable (e.g., they are not outrageous possibilities), the epistemic distinction between them that I have made in the previous paragraph is preserved.[172]

Second, whichever view one takes on the defeasibility of criteria, the important takeaway point is that criteria stand in a unique, conceptual relationship to their corresponding phenomena. We may sometimes be mistaken or have doubts, but repeatedly denying that it is raining when the criteria for RAIN are (or appear to be) met is to make a conceptual error: it is not an exercise of philosophical caution, but arguably a demonstration of the lack of grasp of the concept RAIN.

Anthony Kenny provides a concise summary of the criteria/symptoms distinction, along with his own helpful example:

---

[168] Edward Witherspoon, 'Wittgenstein on Criteria and The Problem Of Other Minds', in *The Oxford Handbook of Wittgenstein* (Oxford University Press, 2011), 484.
[169] Crispin Wright, 'Second Thoughts about Criteria', *Synthese* 58, no. 3 (March 1984): 383–405.
[170] I therefore think it is important to reiterate that the role of criteria is first and foremost conceptual. Here, I agree in part with Jack Temkin: "The point of this story is that the status of some phenomenon as a symptom or as a criterion is not a matter of its evidential strength, but of the centrality of the evidential relation to the meaning of the proposition for which it is evidence." Jack Temkin, 'Wittgenstein on Criteria and Other Minds', *The Southern Journal of Philosophy* 28, no. 4 (1990): 578.
[171] McDowell, *Meaning, Knowledge, and Reality*, 381.
[172] The threat that the possibility of error poses for criteria and symptoms is 'cancelled out', so to speak.

> Where the connection between a certain kind of evidence and the conclusion drawn from it is a matter of empirical discovery, through theory and induction, the evidence may be called a symptom of the state of affairs. Where the relation between evidence and conclusion is not something discovered by empirical investigation, but is something which must be grasped by anyone who possesses the concept of the relevant kind of thing, then the evidence is not a mere symptom, but is a criterion of the state of affairs in question. . . Using this distinction we can say that certain states or events in the brain may be symptoms of certain mental states but they could not be criteria for them in the way that the appropriate behaviour would be. Thus, for instance, certain electrical brain patterns may be, or may some day come to be, symptoms of the possession of a knowledge of English by the person whose brain is in question. But the person's ready use of English is not just a symptom of, it is a criterion for, his knowledge of English.[173]

The AfC finds its motivation from these careful considerations about our concepts and their relationship to our beliefs about the world. The criteria/symptoms distinction also helps to illuminate an important difference between the AA and the ABE on the one hand, and the AfC on the other.[174] Both the AA and the ABE characterize the relationship between consciousness (and various other mental phenomena) and behaviour roughly on the model of the relationship between rain and a falling barometer – i.e., they treat behaviour as just a causal symptom of consciousness to which we are to apply inductive or abductive reasoning. Against this view, the AfC claims that certain behaviour are criteria for consciousness. To see why, let us turn our attention to a Wittgensteinian account of mental concepts.

4.3.2 A Wittgensteinian account of CONSCIOUSNESS

What, on the AfC, is required to grasp mental concepts like CONSCIOUSNESS? Or in its own terminology, what does it say are the criteria for CONSCIOUSNESS? To answer, I will draw upon Wittgenstein's discussions on how we teach mental terms to children.[175] Although these discussions focus on mental *terms*, they contain lessons for mental *concepts*.[176] Note also that,

[173] Anthony Kenny, *The Metaphysics of Mind* (Oxford: Oxford University Press, 1992), 5.
[174] I mean the non-strong ABE (i.e., *not* the version of the ABE discussed in Section 4.2.1).
[175] Wittgenstein, *Philosophical Investigations*, §244.
[176] It is worth noting that Wittgenstein himself (and many of his commentators) often moves freely between talk of mental terms (or words) and mental concepts. This, of course, does not mean that I am right to do the same. However, with the right adjustments, the move seems unproblematic for my purposes. For a discussion of the term/concept distinction in a similar context, see William Child, 'Wittgenstein and Phenomenal Concepts', in *Wittgenstein and Perception* (Routledge, 2015), 86–87.

by drawing upon these discussions, I am not suggesting that grasping mental concepts requires explicit instruction. We may well grasp them through natural exposure (e.g., observing others' usage) or even an innate capacity.[177] As we will see, my claim is only that one's use of mental concepts must be tied up to the natural expressions of mental states and, by extension, to their broader use in our community of concept-users. Relatedly, my (and Wittgenstein's) discussion should not be understood as putting forward a speculative hypothesis about how humans in fact acquire mental concepts. Rather, it is intended only to paint the broad outlines of the conceptual requirements for saying that someone has grasped a mental concept.

Suppose a child comes running to you, crying, grimacing, and clutching her cheek – she seems to have a toothache. According to my Wittgensteinian approach, by teaching the child, "that is what we call a 'toothache'," we are teaching her a kind of new toothache-behaviour. That is, we are teaching her a linguistic and conceptual extension to her natural pre-linguistic toothache-behaviour – a different, more refined way of expressing what she feels. Through repeated lessons of this kind, practice, and correction, as well as general exposure to the numerous circumstances in which we use the concept TOOTHACHE, the child begins to grasp it. As Marie McGinn puts it, "Wittgenstein's example presents us with a sequence in which naturally expressive behaviour becomes progressively more refined and articulate."[178]

On the Wittgensteinian account, then, TOOTHACHE is not grafted, so to speak, onto the sensation of toothache directly, but grafted onto it via its natural behavioural expression.[179] There is little temptation to think here that we grasp TOOTHACHE via introspective attention to toothaches (the phenomenalist account) or via an explanatory theory (the folk-psychological account), for the concept is an extension of our *instinctive* toothache-behaviour. This, as William Child puts it, is "an important corrective to the tendency to overintellectualize our relation to our own minds."[180] The key idea in the AfC lies here: such behaviour are *criteria* for TOOTHACHE. That is, it is through such behaviour that we make sense of and apply the concept. It is in this sense that the argument holds that consciousness (and other mental phenomena) and behaviour are not merely causally or empirically connected, but also *conceptually* connected – we grasp CONSCIOUSNESS through certain characteristic behavioural

[177] Wittgenstein himself does not rule these possibilities out. Wittgenstein, *The Blue and Brown Books*, 12, 96–97. More broadly, this account, as I present it, does not imply that acquiring concepts requires language (e.g., that non-linguistic animals could not possess any concept).
[178] Marie McGinn, *The Routledge Guidebook to Wittgenstein's Philosophical Investigations* (New York: Routledge, 2013), 144.
[179] Anthony Kenny, *Wittgenstein* (Williston, UK: John Wiley & Sons, 2005), 145.
[180] William Child, *Wittgenstein* (Routledge, 2011), 165.

expressions. And it is in this sense that I suggest that certain behaviours are a part of our concept of consciousness. Edward Witherspoon summarizes the idea well:

> When we learn and teach words for inner states, we rely on expressions. To learn how to apply the word "anger" we rely on the typical expressions of anger: we learn that someone who scowls like *that* or who lashes out at a playmate or who yells angrily is angry. This is the basis for a conceptual connection: to know what anger is requires knowing that such behaviors are expressions of anger.[181]

How does the Wittgensteinian account try to capture the phenomenal character of conscious experiences into our mental concepts? It would first deny the phenomenalist idea that grasping mental concepts (or learning to identify mental states) is just a matter of introspective attention to the appropriate mental states. In turn, it denies a conception of conscious experiences as something that cannot be analyzed in terms of other states, including behaviour. To my eyes, the very fact that we can talk and think about conscious experiences renders that idea implausible. At the same time, the Wittgensteinian account does not hold that phenomenal characters are irrelevant to our mental concepts. Its claim is only that, for example, one learns to identify the phenomenal character of pain (that what one is feeling is called 'painfulness') via its connection to certain behavioural patterns and the circumstances in which they take place. The felt character of mental states plays a crucial role in making them what they are, but our ability to pick them out (and form concepts of them) is constrained by behaviour.

There is, then. an important similarity between the Wittgensteinian account and the version of the folk-psychological account introduced at the end of Section 4.2.1. On both accounts, behaviour plays an important role in our grasp of mental concepts. This has the consequence that both accounts reject the asymmetry thesis. But the key difference is that whereas the folk-psychological account draws this connection on theoretical or explanatory grounds, the Wittgensteinian account draws it on conceptual grounds. For example, whereas the folk-psychological account would hold that folk-psychological generalizations such as "people tend to groan when they have a toothache" is an explanatory hypothesis, the Wittgensteinian account would hold that it is simply a conceptual (or 'logico-grammatical') truth (groaning is simply a part of the meaning of 'toothache').

[181] Witherspoon, 'Wittgenstein on Criteria and The Problem Of Other Minds', 487–88. The quote mentions anger (which one may argue is not defined by phenomenal character but is rather more like an attitude or a disposition), but Witherspoon has pain, toothaches, colour perception in mind too.

Moreover, although the folk-psychological account views our understanding of pain as nested in the explanation of behaviour, in doing so, it views our concepts of pain and behaviour as fundamentally distinct – it asks, "what underlies Jones' behaviour?", and the answer is "pain." In fact, in both first- and third-person cases, the same kind of explanatory inference is held to be required. By contrast, the Wittgensteinian view urges us not to view pain as something hidden behind certain behaviour, but as something that is embedded in certain behaviour. The folk-psychological account appeals to an independent phenomenon – pain – to explain behaviour, but on the Wittgensteinian account, behaviour is needed to explain, so to speak, pain. McGinn provides an excellent summary of the idea:

> We don't hear a cry and conjecture that it is accompanied by a particular kind of private object ('THIS'); in appropriate circumstances, we hear a cry of pain, a shriek of fear, a hoot of delight, and so on. The pain, the fear and the delight are not public in the way that the cry, the shriek or the hoot are; but insofar as we experience these sounds as having, in the circumstances, a particular significance, the concepts of pain, fear and delight figure essentially in our description of what we hear.[182]

In Section 4.5, we will see an important consequence of this difference for the problem of AI consciousness. The AA and the ABE (of the folk-psychological variant or not) allow that we can imagine an AI system or alien species that are conscious despite bearing no resemblance to human beings. The AA allows this possibility by virtue of it being an inductive argument which leaves room for alternative routes to its conclusion about another's consciousness. The ABE allows it (as we have seen in Section 3.2) by arguing that a different explanatory theory could, in principle, be devised to explain the behaviour of such AI systems or alien species. By contrast, the AfC which I defend suggests that these views are mistaken.

So far, my focus has been on the matter of the grasp and application of mental concepts in the *first-person*. But before turning to how the Wittgensteinian account captures our relation to *others*, several objections warrant attention. Perhaps the most obvious one, as Wittgenstein anticipated, is that this account amounts to (logical or analytical) behaviourism – that mental concepts are being reduced to behavioural concepts (or mental states to behavioural states).[183] I cannot evaluate behaviourism here, but this assessment is a misunderstanding: the concept

[182] McGinn, *The Routledge Guidebook to Wittgenstein's Philosophical Investigations*, 195–96.
[183] Wittgenstein, *Philosophical Investigations*, §244.

TOOTHACHE (or the linguistic expression 'toothache') replaces, rather than describes or means, natural toothache-behaviour. In other words, TOOTHACHE should be understood as related to toothaches in roughly the way natural toothache-behaviour is – namely, as expressive of them.[184] In turn, the claim is not that toothaches are to be identified *as* toothache-behaviour, but that they can be identified, at bottom, *because of* toothache-behaviour.[185] As Wittgenstein puts it, behavioural criteria help *determine* the meaning of a linguistic expression, rather than *being* its meaning.[186] The following analogy may be helpful: our perceptual experiences of water droplets from the sky help pick out rain (and grasp the concept RAIN) but are not, of course, rain themselves. Hence, although the linguistic expression "I have a toothache" makes sense only because it is associated with criteria such as clutching one's cheek and groaning, the expression no more means "I am clutching my cheeks and groaning" than "it rained yesterday" means "I remember that it rained yesterday." According to the AfC, then, the relationship between behaviour and consciousness is stronger than merely causal, and weaker than identity.

One might complain here against the idea that mental concepts are merely expressions of mental states. Surely an avowal such as "I have a toothache" acts also as a truth-evaluable description of one's toothache? Wittgenstein's own stance on this matter is disputed, but I see no reason to think that such avowals cannot be both expressive and descriptive.[187]

Another objection may point out that one can self-ascribe mental concepts without any particular attention to one's own behaviour. This is true, but no part of the claim that mental concepts are grasped as extensions to natural behavioural responses implies that one must always observe their own behaviour to apply them.[188] Although reference to behaviour is a crucial in our grasp of mental concepts, a competent concept-user will go onto cultivate the ability to self-ascribe mental concepts without relying on behaviour. After all, no part of the claim that some form of acquaintance with water droplets falling from the sky is necessary to grasp RAIN implies that those conditions must always be present to justifiably say that it is

[184] Some have thus characterized Wittgenstein's account of mental concepts as *expressivism*. David H. Finkelstein, *Expression and the Inner* (Harvard University Press, 2003).
[185] Of course, we can identify toothaches through means other than behaviour. This is why I say 'at bottom', to indicate that I am pointing to their conceptual relation – that is, that it is only through natural toothache-behaviour that those other means are made sense of and legitimized.
[186] Ludwig Wittgenstein, *Wittgenstein's Lectures, Cambridge, 1932-35* (Basil Blackwell, 1979), §24.
[187] For a discussion on this matter, see Section 4.3 in Finkelstein, *Expression and the Inner*; Child, *Wittgenstein*, 169–70.
[188] Although the AfC denies the asymmetry thesis as characterized in Section 1.2, it need not deny every form of asymmetry between first- and third-person ascriptions of mental concepts (or knowledge of mental states). Helpful here is Åsa Wikforss's distinction between the strong and the weak asymmetry thesis. Wikforss, 'Knowledge, Belief, and the Asymmetry Thesis'.

raining. That said, it is important to stress that first-person ascriptions of mental concepts must remain consistent with the ways in which they are used by our conceptual community (which at heart stems from the original, natural behavioural expression). If someone self-ascribes TOOTHACHE repeatedly in contexts that conflict with its criteria (e.g., where HEADACHE applies), we should doubt that they have a grasp of TOOTHACHE (and presumably HEADACHE).

Wittgenstein himself anticipates the above concern: "But do I also say in my own case that I am saying something to myself, because I am behaving in such-and-such a way? – I do *not* say it from observation of my behavior. But it only makes sense because I do behave in this way."[189] That is, if I were not the kind of creature that behaves in certain ways in response to certain things, it would not make sense to attribute toothaches (for example) to myself – there would be nowhere for the concept TOOTHACHE to be grafted onto. Again, the Wittgensteinian account's claim is only that we learn that what we feel is what is called 'toothache' through characteristic toothache-behaviour.

Lastly, one might complain that the Wittgensteinian account is circular, for it requires assuming that the child (for example) is conscious – i.e., that the POM is solved. What it needs to show, the objection goes, is precisely how we can be justified in believing that the child *feels* the toothache before we can teach her that what she is experiencing is what we call 'toothache'. The mistake in this objection is that it assumes the phenomenalist account. If we instead identify toothaches as that which is expressed by certain behaviour under certain circumstances as the Wittgensteinian account does, there is no circularity.

Now, what does my Wittgensteinian approach (and, in effect, the AfC) say about the application of mental concepts to others? The first thing to notice is that first- and third-person uses of mental concepts are already intertwined: we teach a child to grasp and make first-person use of TOOTHACHE, for example, initially in circumstances where she expresses natural toothache-behaviour and therefore where we can apply TOOTHACHE to her on that basis.[190] In other words, the grasp and first-person use of TOOTHACHE *presupposes* its intelligible third-person use. At the same time, based on the way the adults have taught her TOOTHACHE in connection to certain behavioural expressions, the child learns that it is in such circumstances that she too can apply the concept to others.

[189] Wittgenstein, *Philosophical Investigations*, §357.
[190] William Child, 'Wittgenstein and Davidson on First-Person Authority and the Univocality of Mental Terms', in *Wittgenstein and Davidson on Language, Thought, and Action*, ed. Claudine Verheggen (Cambridge: Cambridge University Press, 2017), 199–201.

Second, the Wittgensteinian account highlights the role of our natural, instinctive reactions to another's behaviour in the application of mental concepts to others.[191] An intuitive assessment of the phenomenalist and the folk-psychological account seems to suggest that a child's natural (behavioural) reaction to another's toothache-behaviour – say, of distress, avoidance, and concern – is the result of an inductive or abductive inference that they have a toothache. But on the criterial approach, this is to put the cart before the horse: these instinctive reactions to another's toothache-behaviour are not the product but (a part of) the source of our belief that they feel toothache. As Wittgenstein puts it, "Being sure that someone is in pain, doubting whether he is, and so on, are so many natural, instinctive, kinds of behaviour towards other human beings, and our language is merely an auxiliary to, and further extension of, this relation. Our language-game is an extension of primitive behaviour."[192]

Once again, there is little temptation here to think that the child forms a concept of another's toothache from her own toothache-experiences or through a folk-psychological theory, for her reactions to another's toothache-behaviour are as instinctive and natural as her behavioural responses are to her own toothache-experiences. Although there is an element to the third-person application of mental concepts (i.e., natural reactions or attitudes) which is absent (or at least different) in the first-person case, we are taught to make third-person mental attributions on the basis of the same kind of behaviour by which we are taught first-person mental attributions. Altogether, what we have in the Wittgensteinian account of both first- and third-person mental attributions is a story about "a natural pattern of interactions gradually evolving into a verbal interchange."[193]

I should again stress that the circumstance of behaviour plays an important role in learning what certain behavioural patterns are criteria for. For example, what distinguishes pain-behaviour from sorrow-behaviour or a kind smile from a malicious one is their "spatial and temporal context."[194] A child also learns in connection to context that sometimes people exhibit pain-behaviour without pain (e.g., acting on a theatre stage) and sometimes they have pain without exhibiting pain-behaviour (e.g., when they are in an important business meeting). Eventually she will be able to play the same game, so to speak – she may conceal pain to be

[191] Wittgenstein, *Zettel*, §540–545. Thomas Reid seems to present a similar view in the context of the POM. Thomas Reid, *Essays on the Intellectual Powers of Man* (Cambridge Mass: The MIT Press, 1969), chap. 3. For a helpful summary of Reid's account, see Avramides, *Other Minds*, 2000, chap. 7.
[192] Wittgenstein, *Zettel*, §540–545. For a helpful commentary, see Saul A. Kripke, *Wittgenstein on Rules and Private Language: An Elementary Exposition* (Harvard University Press, 1982), 141–142.
[193] Lars Hertzberg, 'Very General Facts of Nature', in *The Oxford Handbook of Wittgenstein*, ed. Oskari Kuusela and Marie McGinn (Oxford University Press, 2011), 368.
[194] Wittgenstein, *Philosophical Investigations*, §539; Wittgenstein, *Zettel*, §492.

complimented as being tough or pretend to be in pain to skip school. As Wittgenstein stresses, although these factors make our use of mental concepts more complicated, they cannot undermine them.[195]

What about mental states that seem to lack uniquely identifiable behavioural expression and cannot be distinguished through context? For example, it is unclear that there are natural behavioural and contextual differences between different (conscious) thoughts or between different perceptual interpretations (e.g., of the Necker cube or the 'duck-rabbit'). It may be that the required public expressions (and criteria) for such states are linguistic (or a very fine-grained dispositional difference).[196]

On the Wittgensteinian picture, then, the basis on which I learn to apply mental concepts in first- and third-person contexts is the same kind of publicly observable behavioural criteria. (This is what Wittgenstein means in his famous remark that "An 'inner process' stands in need of outward criteria."[197]) Observing that others satisfy the criteria for PAIN (or simply, pain), then, allows me to intelligibly attribute pain to them with no concern for whether the same concept that I use for myself is being used for them. Again, a competent concept-user can self-ascribe, say, PAIN without the need to observe their own behaviour, but a grasp of what counts as pain in myself and in others stems from the same set of behavioural criteria. That is, first-person attribution of pain is made within the same conceptual framework as the third-person counterpart. Again, McGinn's summary is helpful:

> The child is not taught to identify private objects that are inside him, but is trained to use language in a way that is essential to our distinctive form of life. He is not taught that others have inside them what he has inside himself, but is trained to notice and to respond, not only to the other's use of language, but to the characteristic patterns of movement, gesture, facial expression, and so on against the background of which our psychological concepts function.[198]

Some authors take the conceptual connection between mental states and behaviour to afford a kind of perceptibility to the mental lives of others.[199] You may recall my brief discussion of

[195] Ludwig Wittgenstein, *Last Writings on the Philosophy of Psychology*, ed. G. H. von Wright and Heikki Nyman, trans. C. G. Luckhardt and Aue, vol. 1 (Oxford: Wiley Blackwell, 1990), §876. He is specifically talking about pain here.
[196] Wittgenstein, *Philosophical Investigations*, §212.
[197] Wittgenstein, §580.
[198] McGinn, *The Routledge Guidebook to Wittgenstein's Philosophical Investigations*, 186.
[199] McDowell, *Meaning, Knowledge, and Reality*; Cassam, 'The Possibility of Knowledge'.

this perceptual account of other minds in Section 1.4. This account accepts that we can, at least sometimes for some mental states, see *that* another creature has them.[200] There are strains in Wittgenstein's writings that lend to this idea and we sometimes do speak as if we have some perceptual access to another's mental states (e.g., "you can see that he is in a lot of pain" or "I like the way you are thinking").[201] The AfC that I am putting forward is compatible with the perceptual account, but I will ultimately remain neutral on this matter.[202]

Before turning to the problem of AI consciousness, there is one more clarification I would like to make. In response to the AfC, Hyslop points out that the fact that "[a] belief is meaningful does not make it justified."[203] The idea, put in terms I have used in this chapter, seems to be that the fact that the application of a concept to others is intelligible does not make it justified. In one sense, this is true: the fact that, for example, the sentence, "it is raining," is intelligible does not entail we are justified in believing that it is raining whenever. But the AfC does not deny this rather trivial truth. More specifically, it is a plain misunderstanding to take the argument to be suggesting that the mere intelligibility of the application of a concept makes it justified. Rather, the AfC's claim is that what makes the application – justified or unjustified, correct or wrong – of, say, the concept PAIN to others intelligible is that there are characteristic behavioural expressions that serve as criteria for the concept. It is when we see that these behavioural expressions are present that we are justified in believing that others are pain. It is not so much, then, that the fact that a belief is intelligible makes it justified, but that behavioural (and circumstantial) criteria, which give the expression of the belief meaning, give justification. Hyslop would be right that the AfC sees an important connection between the intelligibility of our mental concepts (or the meaning of our mental terms) and our knowledge of another's mental states (criteria has a role in both), but the connection is not one of direct entailment.

4.4 AI Consciousness and the AfC

Let us finally turn to what the AfC says about the problem of AI consciousness. How can we justify the belief that an AI system is conscious? According to the AfC, in order for us to deem an AI system conscious, we must carefully consider whether, and to what extent, they satisfy

[200] I should be clear that the idea is usually not that we can see another's mental states but see (the fact) *that* they have them.
[201] Wittgenstein, *Philosophical Investigations*, §225 and II §25.
[202] On this matter, I agree with Gomes that "Claims about the relation of sensations to behaviour don't immediately have implications for the nature and objects of perceptual experiences." Gomes, 'Is There a Problem of Other Minds?', 370.
[203] Hyslop, *Other Minds*, 1995, 73.

our criteria for mental concepts such as CONSCIOUSNESS and PAIN. To do this, I will first examine the more general relationship between the AfC and the problem of AI consciousness, before delving into the AfC's implications for specific AI systems. I will end by considering a challenge from Thomas Nagel that I think is particularly relevant to the problem of AI consciousness.

4.4.1 Must criteria change?

Let me begin with the following remark from Wittgenstein: "only of a living human being and what resembles (behaves like) a living human being can one say: it has sensations; it sees; is blind; hears; is deaf; is conscious or unconscious."[204]

Wittgenstein is not putting forward an analogical-inductive argument here, that since *I* am conscious and human, human behaviour is the gold standard empirical evidence of consciousness in non-human creatures. In line with what I have argued in Section 4.3, he is simply observing that mental concepts such as CONSCIOUSNESS, SENSATION, and SEEING are built on the background of human behaviour and way of life – they find their original home, so to speak, in our appearances and behaviour. But this is not to say that they cannot be extended beyond us – to non-human animals, and, I argue, even to AI systems. Rather, it gives us a blueprint for thinking about the possibility of consciousness in non-human creatures. The AfC's only claim is that the conditions for the use of mental concepts is governed by their original home and that being faithful to this context is a precondition for their intelligible extension elsewhere.

That said, it is interesting to note that Wittgenstein himself seems to set the bar for artificial mentality rather high:

> "Is it possible for a machine to think?" . . . [T]he trouble which is expressed in this question is not really that we don't yet know a machine which could do the job. The question is not analogous to that which someone might have asked a hundred years ago: "Can a machine liquefy a gas?" The trouble is rather that the sentence, "A machine thinks (perceives, wishes)": seems somehow nonsensical. It is as though we had asked "Has the number 3 a colour?"[205]

[204] Wittgenstein, *Philosophical Investigations*, §281.
[205] Wittgenstein, *The Blue and Brown Books*, 47. See also Wittgenstein, *Philosophical Investigations*, §359–360. This particular quote considers the ability to think, perceive, and wish, but the context of his discussion includes other mental phenomena including sensory experience.

As a number of commentators have noted, Wittgenstein seems to be suggesting that non-figuratively attributing mental states to machines involves a conceptual confusion – a category mistake.[206] That is, it is nonsensical, rather than false, to say that a machine has mental states. Some writers use such passages to advance the idea that criteria themselves must morph or expand before we can have any meaningful say in whether or not certain AI systems possess mental states.[207] There is a grain of truth in this idea, for AI systems, at least in the way we currently tend to conceive of them, differ greatly from us at many levels of description. There is something deeply bizarre (rather than outright false) about attributing mental states to AI systems, at least current ones, with a straight face.

Wittgenstein himself is clear that our concepts are subject to change given the shifting of our needs and wants, likening language to an ancient city where new houses, streets, and even whole suburbs can be added.[208] One might thus suggest that how we use mental concepts in the future with respect to AI systems will be a function of the needs and wants of a future society in which they are more widely embedded, and that until then we simply cannot answer the problem of AI consciousness. My suggestion in Section 4.3.2 that our natural reactions to another's behaviour play an important role in learning to apply mental concepts to others also seems to feed into this attitude. Again, there is some useful truth to these readings, for the way we conceive of and relate to machines intuitively matters for our judgements regarding their capacity for consciousness. Arguably, we see something like this with current advancements in AI technology, where the latest chatbots and robots challenge, even change, our deep preconceptions about what AI systems can do and about the nature of our relation to them in a way we did not, and perhaps could not, anticipate. If these suggestions are right, it may be that whether or not AI systems can be deemed conscious is a question reserved for a future where our relation to AI systems is clearer.

---

[206] J. C. Nyiri, 'Wittgenstein and the Problem of Machine Consciousness', *Grazer Philosophische Studien* 33, no. 1 (1989): 385; Stuart G. Shanker, *Ludwig Wittgenstein: Critical Assessments*, vol. 4 (Routledge, 1996), 11; Anthony Kenny, *The Legacy of Wittgenstein* (Blackwell, 1984), 125.

[207] Nyiri, 1989, 385–386. Otto Neumaier takes a similar view for artificial intelligence. Neumaier, 1987.

[208] Wittgenstein, *Philosophical Investigations*, §18; Ludwig Wittgenstein, *Philosophical Grammar*, ed. Rush Rhees, trans. Anthony Kenny, 2005, §69. Elsewhere he suggests that certain aspects of our language use, including those about mental phenomena, are the way they are as a result of a kind of *decision*. Ludwig Wittgenstein, *Last Writings on the Philosophy of Psychology*, ed. Heikki Nyman and G. H. von Wright, vol. 2 (Oxford, UK Cambridge, USA: Wiley-Blackwell, 1993), 8–9. For a useful exposition of this remark in connection to the conceptual problem, see Child, 'Wittgenstein and Davidson on First-Person Authority and the Univocality of Mental Terms'.

I have two thoughts on this conservative outlook. First, it is useful to note that Wittgenstein died in 1951, well before the arrival of modern AI technology and countless other relevant technological advancements. Attributing intelligence, let alone consciousness, to a World War II Enigma machine or the 1948 chess program, *Turochamp*, indeed seems to be a category mistake, but present-day (and near-future) machines are far more advanced and like us than he could have imagined. Accordingly, we should not take Wittgenstein's remark above as closing the door for artificial consciousness, but rather as something of a warning against the hasty anthropomorphization of machines.

Second, we can appreciate the above considerations (even admit the Wittgensteinian flavour in them) without concluding that we are currently in no position to contribute usefully to the problem of AI consciousness. Future AI systems may differ from us in many respects, and our attitude towards and relationship with them is certain to shift and evolve (even in ways we may not have anticipated), but I see little reason to expect such a radically different future that our current criteria for consciousness are inapplicable to them. Moreover, while our criteria may expand and come to encompass new and different kinds of creatures, it seems implausible to me that they will shift in such a way as to include those behaviours that plainly fail to meet our current criteria. In fact, assuming that the kinds of AI systems that will come to be are our decision to make, there is, as far as I can see, little incentive to create AI systems whose behaviours are entirely alien to us.[209]

4.4.2 Replicant, Cog, and LaMDA

Let us now consider what the AfC might say about the three artificial systems introduced in Chapter 3, beginning with Replicant, an imaginary advanced humanoid. Replicant responds in much the same way that Jones does, not just to a piece of Lego, but to countless other kinds of stimuli. His behaviour is consistent and context-sensitive in all the right ways. In my view, we have sufficient reason think that he satisfies our criteria for CONSCIOUSNESS and that we are therefore justified in believing that he is conscious. This is not because he exhibits behaviour which in my own case is caused by consciousness as the AA would suggest (and hence the AfC avoids the one-case and the dubious principle objection). Nor is it because consciousness is the best explanation of his behaviour as the ABE holds (and hence it avoids concerns regarding the unclear explanatory role of consciousness). Instead, it is because his behaviour is expressive of

[209] If there comes a scenario where we lose control of AI advancement, then there will be genuine uncertainty in what we can say. This possibility, however, is largely beyond what I am trying to answer by 'the problem of AI consciousness'.

consciousness, presenting us with precisely the kind of conditions under which we grasp the concept CONSCIOUSNESS. There is no more an inductive or abductive inference from Replicant's pain-behaviour to his pain than there is from observing water droplets falling from a grey, cloudy sky to rain. Replicant's behaviour simply embodies our concept of consciousness.

By holding that consciousness and behaviour are conceptually linked and that we learn to apply CONSCIOUSNESS to ourselves and others through the same behavioural criteria, the AfC avoids the unverifiability objection (according to which the conclusion that Replicant is conscious is logically unverifiable). The AfC simply denies the idea that the consciousness of others is fundamentally hidden: if certain behaviour are the very means by which we learn what it is for a creature (including myself) to be conscious, it cannot fail to be the case that we are warranted in believing that creatures who consistently exhibit such behaviour are conscious. Denying this would be to undermine the basis of the very concept of consciousness and the use that it has in our conceptual community. To borrow Wittgenstein's words, such denial amounts to sawing off the branch one is sitting on.[210] It would be like denying that we can ever truly know that it is raining (or know what rain is) because all we ever observe is the mere appearance of rain. Such a denial contradicts the fact that we grasp the very concept of rain through those appearances. In fact, faced with such a system and observing the complexity of its behaviour first-hand, I believe that any doubt about its status as a conscious being will dissolve away.[211]

That said, care is needed to avoid painting an exaggerated picture of criteria. However much confidence criteria confer on our beliefs, it cannot be beyond that found in our ordinary uses of mental concepts, for those contexts are their very origin. Thus we may at times be mistaken in our judgement of an AI system's consciousness. But we can also mistake the local fire department's waterbombing exercise for rain, and yet we do not seem to take the possibility of error there to undermine our general knowledge of the circumstances of rain – why should we take it to undermine the general claim that AI systems like Replicant are conscious? The possibility of error, doubt, and uncertainty are a natural pattern in the weave of life, as Wittgenstein might put it.[212]

[210] Wittgenstein, *Philosophical Investigations*, §55.
[211] To borrow Wittgenstein's words, in trying to imagine and tell ourselves that Replicant lacks consciousness, we will "find these words becoming quite empty." Wittgenstein, §420. In this passage, Wittgenstein is questioning the claim that we can imagine people around us lacking consciousness despite behaving normally.
[212] Wittgenstein, *Last Writings on the Philosophy of Psychology*, 1990, 1:§862; Wittgenstein, *Zettel*, §568-569. The cited segments concern pretence, but he would say the same for doubt, error, uncertainty, and the like.

The obvious reason one might stress the possibility of error in Replicant's case (and in AI cases more generally) is his unique material basis.[213] Could it be that there is a tacit assumption in the way we come to grasp mental concepts that constrains its application only to the behaviour of carbon-based creatures, ruling out silicon-based AI systems? I find this implausible – as far as I can tell, there is nothing in the Wittgensteinian account I have defended which suggests that there is something unique within *carbon-based* behaviour that gives life, so to speak, to our mental concepts. Suppose that rather than being taught to grasp and use mental concepts by human adults in connection to human behaviour, we are taught them by AI systems in connection to both AI and human behaviour. Such a scenario seems practically possible. Bracketing superficial differences in preferences of food, patterns of rest, and so on owing to our different material composition, I see little reason to think that those concepts somehow fail to refer to genuine mental states.

What might the AfC say about Cog and LaMDA? As I present it, it would agree with both the ABE's conclusion that neither system has any compelling claim to consciousness and its underlying reason that they are too behaviourally simplistic.[214] However, they differ on *why* behavioural simplicity is problematic: the ABE says that Cog and LaMDA's behaviour is poorly explained by consciousness, whereas the AfC would say that their behaviour does not meet our criteria for consciousness.

The challenge for Cog is that it replicates only a very limited range of very basic human behaviour – e.g., shaking hands and turning towards visual and auditory stimuli. We saw in Chapter 3 that Daniel Dennett favourably compares Cog to simple organisms like clams and houseflies, implying that Cog, like they, could be said to have a rudimentary form of consciousness. Although I find that there is some use in thinking of consciousness as coming in varying degrees of complexity (as long as this notion is clearly defined), we saw in Section 3.3.2 that Cog is in fact far simpler than houseflies (and arguably clams) both behaviourally and internally. Further, the differences are not merely quantitative: Cog altogether lacks the mechanisms for detecting and avoiding noxious stimuli. According to the AfC, such types of behaviour play an important part in making our concept of consciousness the concept it is and in undergirding its intelligible application to oneself and to others. Moreover, although Cog possess a face, arms, and fingers, they are not put to the numerous uses we expect of creatures

[213] I have already explained in previous chapters why I find this a poor objection, but I respond to it here in a different manner.

[214] The AA, I think, would conclude the same, on grounds that they resemble only very limited aspects of human behaviour.

that possess those features.[215] Since Cog exhibits only what is effectively a caricature of a fraction of human behaviour, we are not warranted in applying the concept CONSCIOUSNESS (and arguably any mental concept) to it.

I want to pause here to point out that directly observing that a system satisfies the criteria for consciousness is not always necessary to gain the warrant for believing that it is conscious. After all, we are often warranted in saying that it is raining despite being in no position to observe water droplets falling from the sky (e.g., by checking a barometer). In fact, there are circumstances where we do say that someone is conscious despite altogether lacking the capacity for behaviour. For example, total locked-in syndrome patients are believed to possess conscious awareness despite total paralysis. One way to explore the implications of such a case for the problem of AI consciousness is by first considering on what grounds we believe that these patients are conscious. A rigorous examination is beyond my aims here, but some brief comments will be helpful.

One ground is verbal reports from patients who have regained some verbal and/or motor function (in the sort of way that we say someone has dreamt when they tell us so upon waking). Although such recovery is exceptionally rare, successful cases give us a firm ground on which to understand the syndrome.[216] Another possibility is to look for the neural correlates of consciousness or of mental states such as mental imagery and thinking 'yes' or 'no'.[217] Some researchers have used such methods to conclude that there can be 'covert' cognition and even consciousness in a persistent vegetative state traditionally believed to involve wakefulness without conscious awareness.[218] Alternatively, the patient could, in principle at least, be trained to communicate directly via neural codes (rather than using proxies like mental imagery) – e.g., to wilfully activate or modulate their motor or visual cortex to indicate yes-or-no answers.

There are a number of conceptual and methodological hurdles to each of the above methods, but, at least in principle, they provide ways to investigate the relationship between

[215] I think this particular point holds, albeit to a lesser extent, for many present-day robots such as Hanson Robotics' Sophia, Engineered Arts' Ameca, and Boston Dynamic's Atlas.

[216] If Norman Malcolm is right that retrospective dream reports are criteria for dreams, then we might say that retrospective reports from locked-in patients are criteria for the total locked-in syndrome. His view is, however, controversial. Norman Malcolm, 'Dreaming and Skepticism', *The Philosophical Review* 65, no. 1 (1956): 14–37.

[217] Indeed, electroencephalography (EEG) results show near-normal brain function in many locked-in patients. It is also relevant that the primary cause of the syndrome is damage to the lower brain and the brainstem that connects the brain to the body (rather than the cortex which is believed to be more closely implicated in consciousness and cognition). O. N. Markand, 'Electroencephalogram in "Locked-in" Syndrome', *Electroencephalography and Clinical Neurophysiology* 40, no. 5 (May 1976): 529–34.

[218] Owen and Coleman, 'Detecting Awareness in the Vegetative State'; Daniel Kondziella et al., 'Preserved Consciousness in Vegetative and Minimal Conscious States: Systematic Review and Meta-Analysis', *Journal of Neurology, Neurosurgery, and Psychiatry* 87, no. 5 (2016): 485–92.

behaviour and consciousness.[219] Importantly, these methods present no contradiction to the AfC, for they ultimately rely on the empirical correlates of our ordinary behavioural criteria for conscious mental states: voluntary action, verbal report, command-following, and so on. For example, Owen and Coleman's experiments used neural activity associated with imagining playing tennis or navigating one's home as signs of general covert cognition and awareness in persistent vegetative patients.[220] Another trial used the same kinds of activity as signs of yes-or-no answers.[221] From the lens of the AfC, these neural markers are a kind of extensions to our ordinary behavioural criteria for consciousness (and cognition) in much the same way that a falling barometer can be said to be an extension of our criteria for rain. We would be unable to interpret these data *as* evidence of consciousness if they were not explained in connection to such criteria.[222] Nachev and Hacker, writing on total locked-in and persistent vegetative state patients, provide a good summary of this idea: "This is not to say that psychological attributes always have an external manifestation; only that without some link to the outside world, not necessarily a contemporaneous one, they are opaque to discussion, let alone scientific enquiry. . . [W]ithout some manifestation, we can be neither meaningfully contradicted nor supported in our judgements here."[223]

Now, where comparable conditions hold for an AI system, we have reason to suspect that it is conscious despite the lack of relevant behaviour. For instance, if Replicant who in his healthy condition had satisfied our criteria for consciousness suffers an injury that severs or limits the connectivity between his cognitive modules and motor and verbal modules, we might think to investigate whether his artificial neural responses to our probing is indicative of covert consciousness. Now, since Cog and arguably most, if not all, present-day 'bed-ridden' artificial systems do not meet the standards of any of these measures, we have good reason to think that their failure to satisfy our criteria for consciousness does not demand the kind exceptions we give for certain paralyzed human individuals.

Now, what is to be said of LaMDA? The first thing to note is that LaMDA is altogether without a physical body, and its behaviour is confined to generating texts. In my view, the

[219] See, for example, P. Nachev and P. M. S. Hacker, 'Covert Cognition in the Persistent Vegetative State', *Progress in Neurobiology* 91, no. 1 (May 2010): 68–76.
[220] Owen and Coleman, 'Detecting Awareness in the Vegetative State'.
[221] Martin M. Monti et al., 'Willful Modulation of Brain Activity in Disorders of Consciousness', *The New England Journal of Medicine* 362, no. 7 (2010): 579–89.
[222] Of course, by 'behavioural characteristics' I do not mean the paralyzed patient's behaviour, but behaviour associated with consciousness in general – i.e., our criteria for consciousness.
[223] Nachev and Hacker, 'Covert Cognition in the Persistent Vegetative State', 70. See also, Grant Gillett, 'Wittgenstein's Startling Claim: Consciousness and the Persistent Vegetative State', in *Slow Cures and Bad Philosophers* (Duke University Press, 2001), 70–88.

ability to process and produce texts is insufficient to satisfy our criteria for consciousness. It may be tempting to view LaMDA on the model of total locked-in syndrome (or some persistent vegetative state) patients: severely, if not totally, limited in behaviour but exhibiting some evidence of consciousness and cognition. In fact, one might add, LaMDA's responses are far more complex than those of these patients. However, it seems relevant that these human patients are embodied beings that once freely moved about and interacted with the world. They at the very least *used to* unequivocally satisfy our criteria for consciousness and various other physical and cognitive capacities. Moreover, the mechanisms for motor and verbal function are still intact (although severely affected) in these patients, and they arguably possess stored memories of their previous lives as embodied agents. These facts provide some reason to take their neural activity seriously as evidence of covert consciousness.

By contrast, we should be suspicious of the claim that the responses generated by LaMDA, by nature a disembodied system, are indicative of consciousness. It is unclear, for example, whether an inherently disembodied system can imagine playing tennis or navigating through physical space. We should also question whether LaMDA could even come to grasp concepts about our world without any physical or perceptual organ through which to interact with it, and without acquaintance with those features of our life that are home to our concepts (recall my discussion on this matter in Section 3.3.3). Of course, not every concept picks out concrete features of our world – e.g., JUSTICE, LOVE, THREE. But even so, they are invariably tied to our physical world: JUSTICE is tied to (morally weighted) actions, LOVE to interpersonal relationships, and concepts of numbers to measurements of physical objects and patterns. The same goes for *mental* concepts: laying aside the issue whether LaMDA can be said be a subject of mental states, how can it grasp mental concepts without any contact with the behavioural patterns and circumstances that are endemic to them? We thus have little reason to take LaMDA's texts (or any of its other features) as indicative of a mental life.

This is not to say that grasping concepts is as simple as being able to perceive the relevant objects (or relations) and understand that the concepts pick them out – or, more specifically to my purposes, that grasping *mental* concepts is as simple as introspectively attending to the relevant mental states. We have seen that the AfC rejects this phenomenalist picture, suggesting instead that we grasp mental concepts as extensions to certain behavioural patterns and through their many uses in our conceptual/linguistic community. Rather, the point is that it is unclear how a disembodied system can come to stand in the kind of relationship to our world that seems crucial to our ability to grasp concepts. Thus, the comparison between LaMDA (or any chatbot or disembodied system) and locked-in patients is highly dubious.

Even if we allow that LaMDA interacts with and represents the world in its own, unique way, given that its contact with the world consists only of large language databases and, at bottom, abstract programming language and principles, we have little reason to think that the representations are anything like our own or conform to our concepts in any intelligible way (and this seems to hold for future generation of chatbots too). There is certainly some commonality – one could argue, for instance, that our representations or concepts (say, of tennis) are ultimately about the same things.[224] It may also be that the way LaMDA's representations are related to one another is similar to ours (i.e., that its network of representations at least partially mirrors ours). Moreover, not everything about the way a system comes to represent things needs to be identical to that of another system for us to say that we share mutually intelligible representations or concepts. After all, you and I surely do not share an exactly identical understanding of, say, rain (perhaps you have never even seen rain), but we presumably do think and talk about the same phenomenon nonetheless. However, LaMDA's differences to us seem to me to run too deep. Whereas we come to represent (and form concepts of) things through a combination of sensory and perceptual experiences and social interactions, resources available to LaMDA are only, again, large databases of abstract information. They may be said to represent our language, but not thereby our world. (None of the foregoing challenges I have raised for Cog and LaMDA hold for Replicant who is both embodied and behaviourally sophisticated.)

4.4.3 Nagel's challenge

To wrap up this section, let us take a step back from specific AI systems and turn to a challenge to the AfC's application to the problem of AI consciousness. This challenge comes from Thomas Nagel, and although he does not have AI systems in mind, what he says is directly relevant to my claims.[225]

The challenge begins with the suggestion that we can conceive of consciousness in creatures whose behavioural, functional, and physical states differ from ours so much so that they do not provide us with any indication of consciousness as we understand it. In other words, he maintains that a creature can be conscious even though we may not be in a position to identify them as such. He writes:

[224] See, for example, Patrick Butlin, 'Sharing Our Concepts with Machines', *Erkenntnis*, 2021.
[225] Nagel, *The View from Nowhere*; Thomas Nagel, *Mortal Questions* (Cambridge University Press, 1979).

> [We can] think of experiences that we can't imagine. . . The idea is that the concept of mind, though tied to subjectivity, is not restricted to what can be understood in terms of our own subjectivity – what we can translate into the terms of our own experience. We include the subjectively *un*imaginable mental lives of other species, for example, in our conception of the real world without betraying their subjectivity by means of a behaviorist, functionalist, or physicalist reduction. We know there's something there, something perspectival, even if we don't know what it is or even how to think about it.[226]

> There is probably a great deal of life in the universe, and we may be in a position to identify only some of its forms, because we would simply be unable to read as behavior the manifestations of creatures sufficiently unlike us. It certainly means something to speculate that there are such creatures, and that they have minds.[227]

Elsewhere, Nagel makes a similar claim but this time explicitly pitting it against Wittgenstein's contribution to the POM. He correctly describes Wittgenstein's view: "We are left with concepts that are anchored in their application to humans, and that apply to other creatures by a natural extension from the behavioral and contextual criteria that operate in ordinary human cases."[228] However, he continues:

> This seems definitely unsatisfactory, because the experiences of other creatures are certainly independent of the reach of an analogy with the human case. They have their own reality and their own subjectivity. They are not, I assume, of indeterminate character in cases where the natural extension from human behavior and circumstances gives no determinate result. To take a very clear case, if things emerged from a spaceship which we could not be sure were machines or conscious beings, what we were wondering about would have an answer even if the things were so different from anything we were familiar with that we could never discover it. It would depend on whether there was something it was like to be them, not on whether behavioral similarities warranted our saying so. . . So they cannot be analyzed in terms of human criteria for their ascription. And since human experiences have the same kind of reality,

[226] Nagel, 1989, 21.
[227] Nagel, 1989, 24.
[228] Nagel, *Mortal Questions*, 191.

> must not the same be true of them? What they are is not fully captured by an account of the conditions under which first- and third-person ascriptions of experience are appropriate.[229]

The claim that there could be alien species or machines that do not resemble us and yet are conscious is intuitively plausible, even if we cannot put a finger on what such a creature might look like and why we would feel this way. After all, the universe is vast and presumably so is the possible circle of conscious beings. But we should be clear that Nagel, as I read him, does not have in mind any of the creatures that I have discussed such as Cog and LaMDA, or even the fictional characters like Samantha from *Her* and WALL-E. Laying aside the fact that they are not alien species, these systems resemble us in some way, such as seemingly using human language or exhibiting human-like expressions, and these provide grounds for evaluating their status as conscious beings. What kind of creatures, then, does he have in mind?

I see two possible readings. One option is this: there may be creatures out there that are made up of some unknown element, possess radically different internal mechanisms and cognitive architecture, communicate via telepathy, and behave according to goals unknown to us. It would indeed be difficult to determine whether these creatures are conscious using the ordinary behavioural, functional, and physical features associated with human consciousness as a yardstick. However, the problem with this reading is that it seems false that we are without any means to make sense of the possibility that they are conscious. They are clearly an advanced civilization consisting of embodied creatures that possess complex internal mechanisms and exhibit some kind of goal-directed behaviour (including taking the time to visit Earth). Even if we cannot understand exactly what they are doing and why, it seems to me that they do share some very minimal similarities to us – enough for at least some of our mental concepts to latch onto them to some degree.

An alternative reading is that Nagel is interested in the conceivability and possibility of consciousness in creatures whose behavioural, functional, and physical profiles are not *at all* comparable to ours. That is, that the fact that these creatures are conscious cannot be made sense in reference to any of the features associated with human (or terrestrial) consciousness. This seems to me the right reading: these aliens are, Nagel says, "so different from anything we were familiar with that we could never discover [whether they are conscious]."[230] Of course,

[229] Nagel, 191–192.
[230] Nagel, 192.

it seems unlikely that we will design AI systems that are like this, but we can consider Nagel's claim as a general objection to the AfC's distinct conceptual emphasis on human behaviour.

Unfortunately, this view strikes me as not just wildly speculative, but also confused: how can we think of these hypothetical creatures as conscious if, as Nagel himself puts it, their (so-called) consciousness cannot be imagined or analyzed by our criteria for consciousness? (Keep in mind that he believes that we can even *know* that there is "something there, something perspectival.") If the AfC is right, certain characteristic human behaviour – i.e., criteria – are the very means by which we can intelligibly talk and think about consciousness. The use of such criteria is not *betraying*, as Nagel claims, subjectivity – they facilitate our understanding of subjectivity. He applies the concept CONSCIOUSNESS to his alien creatures all the while admitting (at least implicitly) that the conditions for the application of CONSCIOUSNESS are totally absent. I cannot see how this position is coherent, let alone compelling.[231]

It may be tempting to reply thus: "what the aliens have may not be the kind of consciousness as *we* know it, but it may be a kind of consciousness nonetheless." But on what ground can we call what they have 'consciousness'? Whatever they possess, we have no reason to call it 'consciousness' if the criteria for consciousness are not met. It is no coincidence that Nagel does not describe what his imagined aliens may look like – in my view, we cannot even begin to describe such a creature, not because of a lack of imagination, but because it is an unintelligible idea that requires stretching our concept of consciousness beyond recognition. My suspicion is that our reason for thinking that certain alien creatures may be conscious (or intelligent, friendly, dangerous, and so on) is the result of an implicit projection of our criteria for consciousness onto them.

My emphasis on our criteria for consciousness may seem chauvinistic and unduly restrictive on the limits of the distribution of consciousness. Flipping Nagel's example, suppose an alien species visits earth, sees that we are radically different to them in appearance and behaviour, and concludes that (what they call) 'consciousness' does not apply to us. Would we not think that this is unfair – in fact, *wrong*? But we must tread carefully here. If the circumstances that dictate our use of CONSCIOUSNESS do not at all converge with that which dictate the use of (purportedly) the same concept in the alien's life, why should we think that

[231] Nagel himself recognizes the conceptual problem as a serious hurdle for thinking about the distribution of consciousness: According to Nagel, "The interesting problem of other minds is not the epistemological problem. . . It is the conceptual problem, how I can *understand* the attribution of mental states to others. And this in turn is really the problem, how I can conceive of my own mind as merely one of many examples of mental phenomena contained in the world." Nagel, *The View from Nowhere*, 19–20.

the same phenomenon is being considered?[232] (The issue is not a matter of translation since, on my view, successful translation requires commonality in the application of the relevant terms or concepts.) Anthony Kenny's words are helpful here: "If some humanoid creatures used a word which had no such connection with the symptoms and circumstances of pain, it would be difficult to see why the word should be translated as 'pain'."[233] I thus argue that this imagined scenario is confused, for it merely assumes that the aliens share our concept of consciousness.

4.5 Conclusion

Let us retrace our steps. I opened this chapter by explaining the conceptual problem of other minds as a tension between two intuitively plausible claims: (1) grasping CONSCIOUSNESS is matter of introspective attention to one's consciousness, and (2) we intelligibly apply the concept to others. Although few philosophers today take the conceptual problem seriously (or at least take the time to consider it), addressing it is important for setting us up in the right way to tackle the problem of AI consciousness.

Following this suggestion, Section 4.2 placed the AA and the ABE under the scrutiny of the conceptual problem. My view was that there is a plausible way for the ABE to avoid this problem by way of what I named the folk-psychological account of CONSCIOUSNESS. The result, however, is a 'strong' version of the ABE which rejects the asymmetry thesis. By contrast, it was unclear whether a solution is available for the AA. The problem is that it seems built into this argument that there is a fundamental asymmetry between first- and third-person knowledge of consciousness, which raises challenges for the idea that mental concepts like CONSCIOUSNESS and PAIN are applicable to others in the same (if any) sense in which they apply to oneself.

Section 4.3 contained two parts. The first introduced the notion of criteria, distinguishing it from that of symptoms. As I presented it, criteria (with respect to consciousness), at base, are the conditions under which we grasp the concept CONSCIOUSNESS, whereas symptoms are the conditions that are empirically correlated with consciousness (or with our use of the concept CONSCIOUSNESS). The second part saw the development of a Wittgensteinian account of mental concepts on which the argument from criteria (AfC) is built, addressing several objections along the way. The key idea was that our ability to talk and think about consciousness (and other mental phenomena) rests with the fact that our concept of it is tied to behaviour. Although I have used pain as my primary example of a conscious state

[232] See Wittgenstein, *Philosophical Investigations*, §142.
[233] Kenny, *Wittgenstein*, 146. This is Kenny's characterization of Wittgenstein's view.

throughout my discussion, the general Wittgensteinian framework is applicable to many other types of mental phenomena, including thoughts, emotions, and perceptions.

Finally, Section 4.4 brought together the lessons from the previous sections to put forward a justification for believing that certain AI systems are conscious. The key idea here was that it is not on analogical or explanatory grounds that we are justified in believing that a system like Replicant is conscious, but on grounds that his behaviour is the kind of behaviour by which we come to grasp the concept CONSCIOUSNESS in the first place – i.e., on grounds that he satisfies the criteria for CONSCIOUSNESS. I revisited Cog and LaMDA to examine the broader implications of the AfC and concluded that they do not meet these criteria. The section concluded with a counter to Nagel's challenge to Wittgenstein's emphasis on human behaviour.

5. Conclusion

AI technology is advancing at a breakneck speed with no hint of slowing down. Artificial consciousness is no longer just an interesting conversation to have, but also an important one. This thesis has explored one small slice of the many ways this conversation can take place, doing so by borrowing the tools found in the problem of other minds (POM) literature to answer the problem of AI consciousness: how might we justifiably believe that an AI is conscious?

Rather than re-summarize the three arguments I have explored – the argument from analogy (AA), the argument from best explanation (ABE), and the argument from criteria (AfC) – I want to take the time here to reflect on the relationship between these three arguments and raise some of the questions that this thesis leaves open.

I closed Chapter 3 with a comment on the compatibility of the AA and the ABE, not just in terms of directly combining their core elements into a hybrid ABE as we saw in Section 3.2, but also in terms of what Paul Thagard calls joint coherence: where two independent arguments reach the same conclusion in harmonious ways.[234] However, my discussion of the 'strong' version of the ABE and the folk-psychological account of CONSCIOUSNESS in Section 4.2.1 tells me that the compatibility is not so straightforward, for the AA and the strong ABE disagree on the asymmetry thesis. Whereas the AA maintains that first- and third-person

[234] Section 4.4 in Thagard, *Coherence in Thought and Action*.

knowledge of consciousness is distinct both in their source and strength, the strong ABE takes them to be rooted in the same sort of explanatory considerations.[235]

What about the relationship between the AA and the AfC? These two arguments stand in stark contrast – indeed, Wittgenstein's discussions of the POM often explicitly targeted analogical reasoning. At the heart of the disagreement between the two arguments was the AfC's rejection of the idea that the starting point for thinking about the POM (and consciousness more generally) is with our own conscious experiences. That said, I think there is a way of explaining the inclination to project from our own case which the AfC need not reject. Recall Wittgenstein's claim that only of a human being and what behaves like one can we say that it is conscious or unconscious.[236] Although I stressed in Section 4.4.1 that this is not some version of the AA, there is something here that speaks to and helps make sense of our inclination to project from our own case. As I see it, we feel that finding a warrant for the belief that others are conscious is intimately tied to our own case because we come to grasp what it means to be conscious in reference to human behaviour, reaction, and language, and the perspective from which we undergo this lesson is unique to ourselves. This is not to say that there is something inherently superior about first-person knowledge – I am just one of the many unique perspectives. But this seems to me one way of capturing the intuition behind the AA without committing to its problematic conceptual assumptions.

Lastly, the relationship between the ABE and the AfC. In contrast to the AA, Wittgenstein's writings, to my knowledge, do not mention anything like the ABE. This makes sense since the AA was well-known in his time, whereas the ABE really only arrived after his death. But as we pieced his ideas together, the contrast became clear. The main point of contrast is that since the standard ABE (which includes the non-hybrid and the hybrid ABE) accepts the asymmetry thesis, it starts out from a fundamentally different place compared to the AfC. The contrast between the *strong* ABE and the AfC is weaker since both reject the asymmetry thesis, but as I have noted in Section 4.3.2, they do differ in other critical places. In particular, the AfC denies the idea that explanatory theorizing is involved in our grasp of mental concepts and, by extension, in the justification of our belief that others (including certain AI systems) are conscious.

[235] What the strong ABE might say about the apparent difference in the strength of first- and third-person beliefs (i.e., why first-person beliefs seem more reliable) is unclear, but one possible answer which stays true to the argument seems to be to appeal to the richer access one has to their own behaviour (i.e., a greater number of available observations to which explanatory considerations can be applied and found successful).
[236] Wittgenstein, *Philosophical Investigations*, §281.

However, just how we are to understand this disagreement demands further inspection given that the folk-psychological account holds that such explanatory theorizing is an *implicit* or *sub-personal*, rather than conscious, process.[237] Suppose empirical findings strongly suggest that such a process does take place. Would this be incompatible with the AfC? Or would the AfC be able to accommodate those findings without undermining itself? My hunch is that the conceptual robustness of such findings will be questionable from the perspective of the AfC. That said, a number of questions will need to be answered before we can confidently say what the relationship between the two arguments is. For example, does it matter whether the explanatory theorizing takes place at a psychological (i.e., roughly speaking, what the mind does) or a neural (i.e., what the brain does) level? Is describing sub-personal, especially neural, processes as explaining, predicting, processing evidence, and so on, legitimate?[238] Another question concerns how such a sub-personal process can go beyond being an answer to the psychological POM and help us with the epistemic POM.

The AA, the ABE, and the AfC, as I have characterized them, all take behaviour to be the key relevant datum, and they also converge, at least in my analysis, in their conclusions regarding Replicant, Cog, and LaMDA's consciousness.[239] However, because they disagree on the *role* behaviour is to play – inductive, abductive, or criterial/conceptual – the way they derive this conclusion from our observation of behaviour differed. It should also be noted that they may generate conflicting conclusions when it comes to other AI systems.

Although I consider behavioural evidence to be indispensable for answering the problem of AI consciousness, our pre-theoretical intuitions about behaviour require refining. It will be useful here to think about the possible constraints that other levels of description might have on behaviour. It seems unlikely to me that the kinds of behaviour we can implement in an AI system is restricted by the nature of the materials we use, but perhaps causal structure or cognitive architecture is relevant. It will also be helpful here to develop these arguments with a focus on other levels of description (e.g., cognitive architecture, artificial neural network), and see what effect this has on both the plausibility of the arguments as well as their conclusions.

[237] Much like how the information which explains our grasp of the countless rules of language is mostly implicit rather than consciously accessible. Bayne, *Philosophy of Mind*, 200.
[238] One might argue, for example, that this is an instance of the mereological fallacy: "ascribing psychological attributes to parts of an animal that can only intelligibly be ascribed to the animal as a whole." Harry Smit and Peter M. S. Hacker, 'Seven Misconceptions About the Mereological Fallacy: A Compilation for the Perplexed', *Erkenntnis* 79, no. 5 (1 October 2014): 1077. The notion of the mereological fallacy has been partly motivated by Wittgenstein's thoughts, including that expressed in the passage I have examined previously in this thesis. Wittgenstein, *Philosophical Investigations*, §281.
[239] Although, of course, I have not applied the AA directly to Cog and LaMDA.

I have suggested in Section 1.1 that the three solutions are not committed to any particular metaphysical outlook. But they may well dovetail with some better than others. The ABE seems likely to be attractive to functionalists by virtue of its emphasis on the role of mental states (and mental concepts). The AfC may also attract functionalists owing to its emphasis on the ordinary contexts in which we grasp and use mental concepts, but it is also possible that its Wittgensteinian influence attracts behaviourists. By contrast, the AA seems the most metaphysically neutral. Those working on the problem of AI consciousness will thus benefit from keeping their eye on the overall relationship between their project and the metaphysics of consciousness. But they can reach out to other resources as well. For one, interest in the perceptual account of other minds is growing – its discussion now seems to rival that of the ABE both in the psychological (including social cognition) and the epistemic domain.[240] Further, both theory of mind and phenomenology are important neighbouring disciplines to keep in mind.[241] In particular, the phenomenological tradition and the Wittgensteinian roots of the AfC have much in common.

Now, our attribution of consciousness and other kinds of mental states to AI systems will certainly depend in part on the philosophical and scientific state of knowledge. But equally important, to my eyes, is the kind of relationship we have with AI systems. One consideration here is the degree to which they are integrated into our lives. Imagine a future where AI systems do not just resemble us behaviourally but are also integrated into the kind of practices that lie at the heart of human life and our concepts. My view, informed by the AfC, is that in such a world, we will naturally find ourselves attributing conscious deliberation to our humanoid opponents during a bout of chess (although they may not possess a biological brain), pain and distress to a robot pet being mistreated (although they may 'bleed' blue coolant rather than blood), and pleasure and contentment to those robots with whom we form valuable memories (although they may lead very different private lives). Once AI systems develop a rich history of participation in the various aspects of human society and culture, it seems to me we will cease to be tempted to ask whether we are really right to attribute consciousness to them. In other words, it seems to me that the problem of AI consciousness, at least with respect to such AI systems, will slowly but surely dissolve.

[240] Matthew Parrott, 'The Look of Another Mind', *Mind* 126, no. 504 (2017): 1023–61; Anita Avramides, 'Perception, Reliability, and Other Minds', in *Knowing Other Minds*, ed. Anita Avramides and Matthew Parrott (Oxford University Press, 2019), 107–26; Joel Krueger, 'Direct Social Perception', in *The Oxford Handbook of 4E Cognition*, ed. Albert Newen, Leon De Bruin, and Shaun Gallagher (Oxford University Press, 2018), 0; Mason Westfall, 'Other Minds Are Neither Seen nor Inferred', *Synthese* 198, no. 12 (1 December 2021): 11977–97.

[241] Shaun Gallagher Zahavi Dan, *The Phenomenological Mind*, 3rd ed. (London: Routledge, 2020).

This emphasis on the importance of integration has a partly practical purpose, to reduce any biases we might have against attributing consciousness to machines, but for the AfC, it reflects a deeper aspect of our attributional practices: the scaffolding, so to speak, that surrounds our use of mental concepts. It embodies Wittgenstein's remark that our concepts do not merely reflect our life, "[t]hey stand in the middle of it."[242] To think that we can squeeze a wedge between this scaffolding surrounding the phenomenon of consciousness and 'consciousness itself' is a mistake. Of course, there may be AI systems to which we attribute consciousness that are not so integrated – i.e., integration is not necessary. After all, we readily attribute a rich array of mental states to wild animals such as elephants and octopuses. But I think it is still important to point out that our attitude and relationship to AI systems figure into the justification we have for attributing consciousness to them. Since our mental concepts are borne out of our social and cultural life, those AI systems that are integrated into it will make the strongest case as subjects of conscious experiences.

[242] Ludwig Wittgenstein, *Remarks on Colour*, ed. G. E. M. Anscombe, trans. Linda L. McAlister and Margarete Schättle, 2007, §302.

Bibliography

Achinstein, Peter. *The Book of Evidence*. Oxford University Press, 2001.

Addis, Mark R. *Wittgenstein: Making Sense of Other Minds*. Ashgate, 1999.

Aizawa, Kenneth, and Carl Gillett. 'The (Multiple) Realization of Psychological and Other Properties in the Sciences'. *Mind & Language* 24, no. 2 (2009): 181–208. https://doi.org/10.1111/j.1468-0017.2008.01359.x.

Alkire, Michael T., Anthony G. Hudetz, and Giulio Tononi. 'Consciousness and Anesthesia'. *Science (New York, N.Y.)* 322, no. 5903 (2008): 876–80. https://doi.org/10.1126/science.1149213.

Allen, Colin, and Michael Trestman. 'Animal Consciousness'. In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Winter 2020. Metaphysics Research Lab, Stanford University, 2020. https://plato.stanford.edu/archives/win2020/entries/consciousness-animal/.

Augustine. 'De Trinitate'. In *Nicene and Post-Nicene Fathers of the Christian Church*, edited by Philip Schaff, Vol. 3. Grand Rapids: Eerdmans, 1974.

Avramides, A. 'On Seeing That Others Have Thoughts and Feelings'. *Journal of Consciousness Studies* 22, no. 1–2 (2015): 138–55.

Avramides, Anita. *Other Minds*. Routledge, 2000.

———. 'Other Minds'. In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta. Metaphysics Research Lab, Stanford University, 2020. https://plato.stanford.edu/archives/win2020/entries/other-minds/.

———. ‘Perception, Reliability, and Other Minds’. In *Knowing Other Minds*, edited by Anita Avramides and Matthew Parrott, 107–26. Oxford University Press, 2019. https://doi.org/10.1093/oso/9780198794400.003.0006.

Baars, Bernard J. *A Cognitive Theory of Consciousness*. Cambridge University Press, 1988.

Bayne, Tim. *Philosophy of Mind: An Introduction*. New York: Routledge, 2021.

Bayne, Tim, Jakob Hohwy, and Adrian M. Owen. ‘Are There Levels of Consciousness?’ *Trends in Cognitive Sciences* 20, no. 6 (2016): 405–13. https://doi.org/10.1016/j.tics.2016.03.009.

Blackburn, Simon. *Spreading the Word: Groundings in the Philosophy of Language*. Clarendon Press, 1984.

Block, Ned. ‘On a Confusion about a Function of Consciousness’. *Behavioral and Brain Sciences* 18, no. 2 (June 1995): 227–47. https://doi.org/10.1017/S0140525X00038188.

———. ‘The Harder Problem of Consciousness’. *The Journal of Philosophy* 99, no. 8 (2002): 391–425. https://doi.org/10.2307/3655621.

———. ‘Troubles with Functionalism’. *Minnesota Studies in the Philosophy of Science* 9 (1978): 261–325.

Botvinick, Matthew, and Jonathan Cohen. ‘Rubber Hands “Feel” Touch That Eyes See’. *Nature* 391, no. 6669 (1998): 756. https://doi.org/10.1038/35784.

Bringsjord, Selmer, Paul Bello, and David Ferrucci. ‘Creativity, the Turing Test, and the (Better) Lovelace Test’. *Minds and Machines* 11, no. 1 (2001): 3–27. https://doi.org/10.1023/A:1011206622741.

Brooks, Rodney A. ‘Intelligence Without Reason’. In *The Artificial Life Route to Artificial Intelligence*. Routledge, 1995.

Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, et al. ‘Language Models Are Few-Shot Learners’. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, 1877–1901. NIPS’20. Red Hook, NY, USA: Curran Associates Inc., 2020.

Butlin, Patrick. ‘Sharing Our Concepts with Machines’. *Erkenntnis*, 2021. https://doi.org/10.1007/s10670-021-00491-w.

Cassam, Quassim. ‘The Possibility of Knowledge’. In *The Possibility of Knowledge*, edited by Quassim Cassam. Oxford University Press, 2007. https://doi.org/10.1093/acprof:oso/9780199208319.003.0001.

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

Chan, Brenda L., Richard Witt, Alexandra P. Charrow, Amanda Magee, Robin Howard, Paul F. Pasquina, Kenneth M. Heilman, and Jack W. Tsao. 'Mirror Therapy for Phantom Limb Pain'. *The New England Journal of Medicine* 357, no. 21 (2007): 2206–7. https://doi.org/10.1056/NEJMc071927.

Chihara, Charles S., and Jerry A. Fodor. 'Operationalism and Ordinary Language: A Critique of Wittgenstein'. *American Philosophical Quarterly* 2, no. 4 (1965): 281–95.

Child, William. *Wittgenstein*. Routledge, 2011.

———. 'Wittgenstein and Davidson on First-Person Authority and the Univocality of Mental Terms'. In *Wittgenstein and Davidson on Language, Thought, and Action*, edited by Claudine Verheggen, 159–85. Cambridge: Cambridge University Press, 2017. https://doi.org/10.1017/9781316145364.010.

———. 'Wittgenstein and Phenomenal Concepts'. In *Wittgenstein and Perception*. Routledge, 2015.

Churchland, Paul M. *Matter and Consciousness*. The MIT Press. The MIT Press, 2013.

Cochrane, Tom. 'A Case of Shared Consciousness'. *Synthese* 199, no. 1 (2021): 1019–37. https://doi.org/10.1007/s11229-020-02753-6.

Cottingham, John, Robert Stoothoff, and Dugald Murdoch. *The Philosophical Writings of Descartes: Volume 2*. Cambridge University Press, 1984.

Crane, Tim. 'The Origins of Qualia'. In *The History of the Mind-Body Problem*, edited by Tim Crane and Sarah Patterson. London: Routledge, 2000.

Davidson, Donald. *Subjective, Intersubjective, Objective: Philosophical Essays Volume 3*. Clarendon Press, 2001.

Dehaene, Stanislas, Hakwan Lau, and Sid Kouider. 'What Is Consciousness, and Could Machines Have It?', 2017, 8.

Dennett, Daniel. 'Intentional Systems Theory'. In *The Oxford Handbook of Philosophy of Mind*, edited by Ansgar Beckermann, Brian P. McLaughlin, and Sven Walter, 1st ed., 339–50. Oxford University Press, 2009. https://doi.org/10.1093/oxfordhb/9780199262618.003.0020.

———. 'Review of Other Minds: The Octopus, the Sea and the Deep Origins of Consciousness.' *Biology & Philosophy* 34, no. 1 (2018).

Dennett, Daniel C. 'Cog: Steps Toward Consciousness in Robots'. In *Conscious Experience*, edited by Thomas Metzinger, 471–87. Ferdinand Schoningh, 1995.

Dennett, Daniel C., F. Dretske, S. Shurville, A. Clark, I. Aleksander, and J. Cornwell. 'The Practical Requirements for Making a Conscious Robot [and Discussion]'. *Philosophical Transactions: Physical Sciences and Engineering* 349, no. 1689 (1994): 133–46.

Earl, Brian. 'The Biological Function of Consciousness'. *Frontiers in Psychology* 5 (5 August 2014): 697. https://doi.org/10.3389/fpsyg.2014.00697.

Evenden, Ian. 'What Is ChatGPT? The AI Chatbot Explained'. *Stuff* (blog), 8 February 2023. https://www.stuff.tv/features/what-is-chatgpt-the-ai-chatbot-explained/.

Finkelstein, David H. *Expression and the Inner*. Harvard University Press, 2003.

Fodor, Jerry A. *Psychosemantics: The Problem of Meaning in the Philosophy of Mind*. MIT Press, 1987.

French, Robert M. 'Subcognition and the Limits of the Turing Test'. *Mind* 99, no. 393 (1990): 53–65.

Gamez, David. *Human and Machine Consciousness*. Open Book Publishers, 2018. https://doi.org/10.11647/OBP.0107.

———. *Human and Machine Consciousness*. Open Book Publishers, 2018. https://doi.org/10.11647/obp.0107.

Gillett, Grant. 'Wittgenstein's Startling Claim: Consciousness and the Persistent Vegetative State'. In *Slow Cures and Bad Philosophers*, 70–88. Duke University Press, 2001. https://doi.org/10.1515/9780822381266-007.

Godfrey-Smith, Peter. *Other Minds: The Octopus, the Sea, and the Deep Origins of Consciousness*. First edition. New York: Farrar, Straus and Giroux, 2016.

Gomes, Anil. 'Is There a Problem of Other Minds?' *Proceedings of the Aristotelian Society* 111 (2011): 353–73.

———. 'Skepticism About Other Minds'. In *Skepticism: From Antiquity to the Present*, edited by Diego Machuca and Baron Reed, 700–713. Bloomsbury Academic, 2018.

Guttenplan, Samuel. *A Companion to the Philosophy of Mind*. Cambridge: Blackwell, 1994.

Hacker, P. M. S. *Wittgenstein: On Human Nature*. Phoenix, 1997.

Halgren, E. 'Mental Phenomena Induced by Stimulation in the Limbic System'. *Human Neurobiology* 1, no. 4 (1982): 251–60.

Harré, Rom. 'Wittgenstein and Artificial Intelligence'. *Philosophical Psychology* 1, no. 1 (1988): 105–15. https://doi.org/10.1080/09515088808572928.

Held, Richard, Yuri Ostrovsky, Beatrice de Gelder, Tapan Gandhi, Suma Ganesh, Umang Mathur, and Pawan Sinha. 'The Newly Sighted Fail to Match Seen with Felt'. *Nature Neuroscience* 14, no. 5 (2011): 551–53. https://doi.org/10.1038/nn.2795.

Hertzberg, Lars. 'Very General Facts of Nature'. In *The Oxford Handbook of Wittgenstein*, edited by Oskari Kuusela and Marie McGinn, 0. Oxford University Press, 2011. https://doi.org/10.1093/oxfordhb/9780199287505.003.0017.

Hirstein, William. 'Mindmelding: Connected Brains and the Problem of Consciousness'. *Mens Sana Monographs* 6, no. 1 (2008): 110–30. https://doi.org/10.4103/0973-1229.38516.

———. 'Sharing Conscious States'. In *Mindmelding: Consciousness, Neuroscience, and the Mind's Privacy*, edited by William Hirstein, 148–64. Oxford University Press, 2012. https://doi.org/10.1093/acprof:oso/9780199231904.003.0009.

Ho, Manh-Tung. 'What Is a Turing Test for Emotional AI?' *AI & SOCIETY*, 2022. https://doi.org/10.1007/s00146-022-01571-3.

Hohwy, Jakob, and Tim Bayne. 'Causes, Confounds and Constituents: The Neural Correlates of Consciousness'. In *The Constitution of Phenomenal Consciousness: Toward a Science and Theory*, edited by Steven M. Miller, 155–76. Advances in Consciousness Research. John Benjamins Publishing Company, 2015. https://doi.org/10.1075/aicr.92.06hoh.

Hutto, Daniel, and Ian Ravenscroft. 'Folk Psychology as a Theory'. In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Fall 2021. Metaphysics Research Lab, Stanford University, 2021. https://plato.stanford.edu/archives/fall2021/entries/folkpsych-theory/.

Hyslop, A., and F. C. Jackson. 'The Analogical Inference to Other Minds'. *American Philosophical Quarterly* 9, no. 2 (1972): 168–76.

Hyslop, Alec. *Other Minds*. Synthese Library, v. 246. Dordrecht ; Boston: Kluwer Academic Publishers, 1995.

———. 'Other Minds'. In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta,. Metaphysics Research Lab, Stanford University, 2014. https://plato.stanford.edu/archives/spr2019/entries/other-minds/.

———. 'Other Minds as Theoretical Entities'. *Australasian Journal of Philosophy* 54, no. 2 (1 August 1976): 158–61. https://doi.org/10.1080/00048407612341171.

Jollimore, Troy. '"This Endless Space between the Words": The Limits of Love in Spike Jonze's Her'. *Midwest Studies In Philosophy* 39, no. 1 (2015): 120–43. https://doi.org/10.1111/misp.12039.

Kenny, Anthony. *The Legacy of Wittgenstein*. Blackwell, 1984.

———. *The Metaphysics of Mind*. Oxford: Oxford University Press, 1992. https://doi.org/10.1093/acprof:oso/9780192830708.001.0001.

———. *Wittgenstein*. Williston, UK: John Wiley & Sons, 2005.

Kirk, Robert. *Robots, Zombies and Us: Understanding Consciousness*. New York: Bloomsbury Academic, 2017.

Kondziella, Daniel, Christian K. Friberg, Vibe G. Frokjaer, Martin Fabricius, and Kirsten Møller. 'Preserved Consciousness in Vegetative and Minimal Conscious States: Systematic Review and Meta-Analysis'. *Journal of Neurology, Neurosurgery, and Psychiatry* 87, no. 5 (2016): 485–92. https://doi.org/10.1136/jnnp-2015-310958.

Kripke, Saul A. *Wittgenstein on Rules and Private Language: An Elementary Exposition*. Harvard University Press, 1982.

Krueger, Joel. 'Direct Social Perception'. In *The Oxford Handbook of 4E Cognition*, edited by Albert Newen, Leon De Bruin, and Shaun Gallagher, 0. Oxford University Press, 2018. https://doi.org/10.1093/oxfordhb/9780198735410.013.15.

Krueger, Joel, and Søren Overgaard. 'Seeing Subjectivity: Defending a Perceptual Account of Other Minds'. In *Seeing Subjectivity: Defending a Perceptual Account of Other Minds*, 297–320. De Gruyter, 2013. https://doi.org/10.1515/9783110325843.297.

Kumar Sharma, Ramesh. 'Dharmakīrti on the Existence of Other Minds'. *Journal of Indian Philosophy* 13, no. 1 (1 March 1985): 55–71. https://doi.org/10.1007/BF00208527.

Lagercrantz, Hugo, and Jean-Pierre Changeux. 'The Emergence of Human Consciousness: From Fetal to Neonatal Life'. *Pediatric Research* 65, no. 3 (March 2009): 255–60. https://doi.org/10.1203/PDR.0b013e3181973b0d.

Google. 'LaMDA: Our Breakthrough Conversation Technology', 18 May 2021. https://blog.google/technology/ai/lamda/.

Lavelle, Jane Suilin. *The Social Mind: A Philosophical Introduction*. London: Routledge, 2018. https://doi.org/10.4324/9781315735535.

Lemoine, Blake. 'Is LaMDA Sentient? — An Interview'. *Medium* (blog), 11 June 2022. https://cajundiscordian.medium.com/is-lamda-sentient-an-interview-ea64d916d917.

Levy, Steven. ‘Blake Lemoine Says Google’s LaMDA AI Faces “Bigotry”’. *Wired*. Accessed 14 November 2022. https://www.wired.com/story/blake-lemoine-google-lamda-ai-bigotry/.

Lipton, Peter. *Inference to the Best Explanation*. 2nd ed. London: Routledge, 2004. https://doi.org/10.4324/9780203470855.

Lloyd, Seth. ‘A Turing Test for Free Will’. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 370, no. 1971 (2012): 3597–3610. https://doi.org/10.1098/rsta.2011.0331.

Locke, Don. ‘Just What Is Wrong with the Argument from Analogy?’ *Australasian Journal of Philosophy* 51, no. 2 (1973): 153–56. https://doi.org/10.1080/00048407312341171.

Lycan, William. ‘Representational Theories of Consciousness’. In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Fall 2019. Metaphysics Research Lab, Stanford University, 2019. https://plato.stanford.edu/archives/fall2019/entries/consciousness-representational/.

Malcolm, Norman. ‘Dreaming and Skepticism’. *The Philosophical Review* 65, no. 1 (1956): 14–37. https://doi.org/10.2307/2182186.

———. ‘I. Knowledge of Other Minds’. *The Journal of Philosophy* 55, no. 23 (1958): 969–78. https://doi.org/10.2307/2021905.

Markand, O. N. ‘Electroencephalogram in “Locked-in” Syndrome’. *Electroencephalography and Clinical Neurophysiology* 40, no. 5 (May 1976): 529–34. https://doi.org/10.1016/0013-4694(76)90083-3.

McDowell, John. *Meaning, Knowledge, and Reality*. Cambridge, MA: Harvard University Press, 1998.

McGinn, Marie. *The Routledge Guidebook to Wittgenstein’s Philosophical Investigations*. New York: Routledge, 2013.

Mckilliam, Andy Kenneth. ‘What Is a Global State of Consciousness?’ *Philosophy and the Mind Sciences* 1, no. 2 (2020). https://doi.org/10.33735/phimisci.2020.II.58.

McLaughlin, Brian P. ‘A Naturalist-Phenomenal Realist Response to Block’s Harder Problem’. *Philosophical Issues* 13 (2003): 163–204.

Melnyk, Andrew. ‘Inference to the Best Explanation and Other Minds’. *Australasian Journal of Philosophy* 72, no. 4 (1994): 482–91. https://doi.org/10.1080/00048409412346281.

Metzinger, Thomas. ‘Artificial Suffering: An Argument for a Global Moratorium on Synthetic Phenomenology’. *Journal of Artificial Intelligence and Consciousness* 8, no. 1 (2021): 43–66. https://doi.org/10.1142/S270507852150003X.

———. *Being No One: The Self-Model Theory of Subjectivity*. MIT Press, 2003.

Mill, John Stuart. *An Examination of Sir William Hamilton's Philosophy: Volume 9*. Vol. 9. University of Toronto Press, 1979.

Mole, Christopher. 'Consciousness and Attention'. In *The Oxford Handbook of the Philosophy of Consciousness*, edited by Uriah Kriegel, 0. Oxford University Press, 2020. https://doi.org/10.1093/oxfordhb/9780198749677.013.23.

Monti, Martin M., Audrey Vanhaudenhuyse, Martin R. Coleman, Melanie Boly, John D. Pickard, Luaba Tshibanda, Adrian M. Owen, and Steven Laureys. 'Willful Modulation of Brain Activity in Disorders of Consciousness'. *The New England Journal of Medicine* 362, no. 7 (2010): 579–89. https://doi.org/10.1056/NEJMoa0905370.

Nachev, P., and P. M. S. Hacker. 'Covert Cognition in the Persistent Vegetative State'. *Progress in Neurobiology* 91, no. 1 (2010): 68–76. https://doi.org/10.1016/j.pneurobio.2010.01.009.

Nachev, Parashkev, Christopher Kennard, and Masud Husain. 'Functional Role of the Supplementary and Pre-Supplementary Motor Areas'. *Nature Reviews Neuroscience* 9, no. 11 (November 2008): 856–69. https://doi.org/10.1038/nrn2478.

Nagel, Thomas. *Mortal Questions*. Cambridge University Press, 1979.

———. *The View from Nowhere*. New York, United States: Oxford University Press, 1989.

———. 'What Is It Like to Be a Bat?' *The Philosophical Review* 83, no. 4 (1974): 435–50. https://doi.org/10.2307/2183914.

Noë, Alva. *Out of Our Heads: Why You Are Not Your Brain, and Other Lessons From the Biology of Consciousness*. Hill & Wang, 2009.

Nyiri, J. C. 'Wittgenstein and the Problem of Machine Consciousness'. *Grazer Philosophische Studien* 33, no. 1 (1989): 375–94. https://doi.org/10.1163/18756735-90000405.

Overgaard, Morten. 'Insect Consciousness'. *Frontiers in Behavioral Neuroscience* 15 (2021). https://doi.org/10.3389/fnbeh.2021.653041.

Owen, Adrian M., and Martin R. Coleman. 'Detecting Awareness in the Vegetative State'. *Annals of the New York Academy of Sciences* 1129 (2008): 130–38. https://doi.org/10.1196/annals.1417.018.

Pargetter, Robert. 'The Scientific Inference to Other Minds'. *Australasian Journal of Philosophy* 62, no. 2 (1984): 158–63. https://doi.org/10.1080/00048408412341341.

Parrott, Matthew. ‘Enquiries Concerning the Minds of Others’. In *Knowing Other Minds*, edited by Anita Avramides and Matthew Parrott, 0. Oxford University Press, 2019. https://doi.org/10.1093/oso/9780198794400.003.0001.

———. ‘The Look of Another Mind’. *Mind* 126, no. 504 (2017): 1023–61. https://doi.org/10.1093/mind/fzw001.

Peacocke, Christopher, and Colin McGinn. ‘Consciousness and Other Minds’. *Proceedings of the Aristotelian Society, Supplementary Volumes* 58 (1984): 97–137.

Proudfoot, Diane. ‘The Turing Test—from Every Angle’. In *The Turing Guide*, edited by Jack Copeland, Jonathan Bowen, Mark Sprevak, and Robin Wilson, 287–300. Oxford University Press, 2017. https://doi.org/10.1093/oso/9780198747826.003.0037.

Putnam, Hilary, ed. ‘Other Minds’. In *Philosophical Papers: Volume 2: Mind, Language and Reality*, 2:342–61. Cambridge: Cambridge University Press, 1975. https://doi.org/10.1017/CBO9780511625251.019.

Raji, Joshua I., and Christopher J. Potter. ‘The Number of Neurons in Drosophila and Mosquito Brains’. *PLOS ONE* 16, no. 5 (2021): e0250381. https://doi.org/10.1371/journal.pone.0250381.

Reggia, James A. ‘The Rise of Machine Consciousness: Studying Consciousness with Computational Models’. *Neural Networks* 44 (2013): 112–31. https://doi.org/10.1016/j.neunet.2013.03.011.

Reggia, James A., Garrett E. Katz, and Gregory P. Davis. ‘Artificial Conscious Intelligence’. *Journal of Artificial Intelligence and Consciousness* 7, no. 1 (2020): 95–107. https://doi.org/10.1142/S270507852050006X.

Reid, Thomas. *Essays on the Intellectual Powers of Man*. Cambridge Mass: The MIT Press, 1969.

Robbins, Philip, and Anthony I. Jack. ‘The Phenomenal Stance’. *Philosophical Studies* 127, no. 1 (1 January 2006): 59–85. https://doi.org/10.1007/s11098-005-1730-x.

Schindler, Samuel. ‘Theoretical Virtues: Do Scientists Think What Philosophers Think They Ought to Think?’ *Philosophy of Science* 89, no. 3 (2022): 542–64. https://doi.org/10.1017/psa.2021.40.

Searle, John. *The Rediscovery of the Mind*. MIT Press, 1992.

Seth, Anil. *Being You: A New Science of Consciousness*. New York, New York: Dutton, 2021.

———. ‘The Strength of Weak Artificial Consciousness’. *International Journal of Machine Consciousness* 1, no. 1 (2009): 71–82. https://doi.org/10.1142/S1793843009000086.

Shah, Huma, and Kevin Warwick. ‘Machine Humour: Examples From Turing Test Experiments’. *AI and Society* 32, no. 4 (2017): 553–61. https://doi.org/10.1007/s00146-016-0669-0.

Shanker, Stuart G. *Ludwig Wittgenstein: Critical Assessments*. Vol. 4. Routledge, 1996.

Slater, Mel, Bernhard Spanlang, Maria V. Sanchez-Vives, and Olaf Blanke. ‘First Person Experience of Body Transfer in Virtual Reality’. *PLOS ONE* 5, no. 5 (2010): e10564. https://doi.org/10.1371/journal.pone.0010564.

Slors, Marc V. P. ‘Intentional Systems Theory, Mental Causation and Empathic Resonance’. *Erkenntnis* 67, no. 2 (2007): 321–36. https://doi.org/10.1007/s10670-007-9074-x.

Smit, Harry, and Peter M. S. Hacker. ‘Seven Misconceptions About the Mereological Fallacy: A Compilation for the Perplexed’. *Erkenntnis* 79, no. 5 (1 October 2014): 1077–97. https://doi.org/10.1007/s10670-013-9594-5.

Sparrow, Robert. ‘Can Machines Be People? Reflections on the Turing Triage Test’. *Robot Ethics: The Ethical and Social Implications of Robotics*, 2012, 301–15.

Stemmer, Nathan. ‘The Hypothesis of Other Minds: Is It the Best Explanation?’ *Philosophical Studies* 51, no. 1 (1987): 109–21. https://doi.org/10.1007/bf00353966.

Tantiwisawaruji, Sukanlaya, Maria J. Rocha, Ana Silva, Miguel A. Pardal, Uthaiwan Kovitvadhi, and Eduardo Rocha. ‘A Stereological Study of the Three Types of Ganglia of Male, Female, and Undifferentiated Scrobicularia Plana (Bivalvia)’. *Animals* 12, no. 17 (2022): 2248. https://doi.org/10.3390/ani12172248.

Temkin, Jack. ‘Wittgenstein on Criteria and Other Minds’. *The Southern Journal of Philosophy* 28, no. 4 (1990): 561–93. https://doi.org/10.1111/j.2041-6962.1990.tb00559.x.

Thagard, Paul. *Coherence in Thought and Action*. MIT Press, 2002.

‘The Secret of Consciousness, with Daniel C. Dennett | New Philosopher’. Accessed 20 February 2023. https://www.newphilosopher.com/articles/the-secret-of-consciousness-with-daniel-c-dennett/.

Thomas, Janice. ‘Mill’s Argument for Other Minds’. *British Journal for the History of Philosophy* 9, no. 3 (2001): 507–23. https://doi.org/10.1080/09608780110072498.

Tiku, Nitasha. ‘The Google Engineer Who Thinks the Company’s AI Has Come to Life’. *Washington Post*, 17 June 2022. https://www.washingtonpost.com/technology/2022/06/11/google-ai-lamda-blake-lemoine/.

Tononi, Giulio, and Christof Koch. ‘Consciousness: Here, There and Everywhere?’ *Philosophical Transactions of the Royal Society B: Biological Sciences* 370, no. 1668 (2015): 20140167. https://doi.org/10.1098/rstb.2014.0167.

Turing, A. M. ‘Computing Machinery and Intelligence’. *Mind*, 1950, 433–60. https://doi.org/10.1093/mind/LIX.236.433.

Turing, Alan. ‘Intelligent Machinery (1948)’, 2004. https://doi.org/10.1093/oso/9780198250791.003.0016.

———. ‘Intelligent Machinery (1948)’. In *The Essential Turing*, edited by B J Copeland, 0. Oxford University Press, 2004. https://doi.org/10.1093/oso/9780198250791.003.0016.

Tye, Michael. *Consciousness Revisited: Materialism Without Phenomenal Concepts*. MIT Press, 2008.

———. *Tense Bees and Shell-Shocked Crabs: Are Animals Conscious?* Oxford University Press USA, 2016.

Van Gulick, Robert. ‘Consciousness’. In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Winter 2021. Metaphysics Research Lab, Stanford University, 2021. https://plato.stanford.edu/archives/win2021/entries/consciousness/.

Waterman, Craig M. ‘The Turing Test and the Argument from Analogy for Other Minds’. *Southwest Philosophy Review* 11, no. 1 (1995): 15–22. https://doi.org/10.5840/swphilreview19951112.

Wegner, Daniel M. *The Illusion of Conscious Will*. Cambridge, Massachusetts: MIT Press, 2002.

Wellman, Carl. ‘Our Criteria for Third-Person Psychological Sentences’. *The Journal of Philosophy* 58, no. 11 (1961): 281–93. https://doi.org/10.2307/2023623.

Westfall, Mason. ‘Other Minds Are Neither Seen nor Inferred’. *Synthese* 198, no. 12 (1 December 2021): 11977–97. https://doi.org/10.1007/s11229-020-02844-4.

Wikforss, Åsa. ‘Knowledge, Belief, and the Asymmetry Thesis’. In *Knowing Other Minds*. Oxford: Oxford University Press, 2019. https://doi.org/10.1093/oso/9780198794400.003.0003.

Windt, Jennifer M., Tore Nielsen, and Evan Thompson. ‘Does Consciousness Disappear in Dreamless Sleep?’ *Trends in Cognitive Sciences* 20, no. 12 (2016): 871–82. https://doi.org/10.1016/j.tics.2016.09.006.

Witherspoon, Edward. ‘Wittgenstein on Criteria and The Problem Of Other Minds’. In *The Oxford Handbook of Wittgenstein*. Oxford University Press, 2011. https://doi.org/10.1093/oxfordhb/9780199287505.003.0022.

Wittgenstein, Ludwig. *Last Writings on the Philosophy of Psychology*. Edited by G. H. von Wright and Heikki Nyman. Translated by C. G. Luckhardt and Aue. Vol. 1. Oxford: Wiley Blackwell, 1990.

———. *Last Writings on the Philosophy of Psychology*. Edited by Heikki Nyman and G. H. von Wright. Vol. 2. Oxford, UK Cambridge, USA: Wiley-Blackwell, 1993.

———. *Philosophical Grammar*. Edited by Rush Rhees. Translated by Anthony Kenny, 2005.

———. *Philosophical Investigations*. Edited by P. M. S. Hacker and Joachim Schulte. 4th Edition. Wiley-Blackwell, 2009.

———. *Remarks on Colour*. Edited by G. E. M. Anscombe. Translated by Linda L. McAlister and Margarete Schättle, 2007.

———. *The Blue and Brown Books: Preliminary Studies for the 'Philosophical Investigations'*. Oxford, England: Harper & Row, 1958.

———. *Wittgenstein's Lectures, Cambridge, 1932-35*. Basil Blackwell, 1979.

———. *Zettel*. Berkeley and Los Angeles: Blackwell, 1967.

Wright, Crispin. 'Second Thoughts about Criteria'. *Synthese* 58, no. 3 (March 1984): 383–405. https://doi.org/10.1007/BF00485248.

Zahavi, Shaun Gallagher, Dan. *The Phenomenological Mind*. 3rd ed. London: Routledge, 2020. https://doi.org/10.4324/9780429319792.