**The hard problem of consciousness and panpsychism: investigating a foundation for artificial general intelligence**

by

Willem Hendrik Louw

submitted in accordance with the requirements
for the degree of

MASTER OF ARTS

in the subject

Philosophy

at the

UNIVERSITY OF SOUTH AFRICA

SUPERVISOR: Prof. Callum David Scott

CO-SUPERVISOR: Ms Sanelisiwe Primrose Ndlovu

2026

## Declaration

**Student Number:** 62296205

I, Willem Hendrik Louw, affirm that this dissertation **- The hard problem of consciousness and panpsychism: investigating a foundation for artificial general intelligence –** is my own original work and that all sources referenced or quoted have been acknowledged and cited in accordance with University of South Africa guidelines.

**2026/02/24**

**W.H. Louw**

## Abstract

The field of artificial intelligence has historically proceeded on computationalist accounts of the mind, which are a subset of physicalist accounts of the mind, that ignore metaphysical issues of consciousness posed by assuming physicalism. However, as the technology advances, the distinction between simulating behaviour and instantiating real understanding becomes critical. This study investigates the ontological foundations of artificial general intelligence (AGI), arguing that the prevailing physicalist and functionalist frameworks fail to give a complete account of consciousness. David Chalmers' Hard Problem of Consciousness is analysed to demonstrate that a purely functional or structural description of the mind leaves an explanatory gap regarding subjective experience. Responding to this ontological challenge, the study proposes panpsychism as a viable alternate foundation for AGI. By grounding phenomenal properties at the fundamental level as intrinsic to matter, panpsychism resolves the problem of brute emergence that plagues physicalism. It is argued that this metaphysical reframing has profound implications for the AGI project. Ultimately, such a shift moves the goal of AGI further away, without denying it outright. The study concludes that panpsychism is a necessary ontological foundation for AGI, wherein consciousness is understood as an event of participation and not merely a physical process.

**Key terms:** Philosophy of Mind; Artificial Intelligence (AI); Artificial General Intelligence (AGI); Metaphysics; Consciousness; Hard problem of consciousness; Philosophy of Science; Machine Consciousness; Panpsychism; Russellian Monism

## Opsomming (Abstract: Afrikaans)

Die veld van kunsmatige intelligensie het histories die fisikalistiese aanname aanvaar dat intelligensie rekenkundig nageboots kan word, onderwyl dit die filosofiese probleem van bewussyn wat met die aanname gepaard gaan grootliks geïgnoreer het. Namate die tegnologie ontwikkel, raak die verskil tussen die nabootsing van gedrag en die werklike vestiging van begrip van kritieke belang. Hierdie studie ondersoek die ontologiese grondslag van kunsmatige algemene intelligensie (KAI/AGI) en voer aan dat die heersende fisikalistiese en funksionalistiese raamwerke nie ’n volledige verduideliking van bewussyn gee nie. David Chalmers se Moeilike Probleem van Bewussyn word ontleed om te demonstreer dat ’n suiwer funksionele of strukturele beskrywing van die verstand ’n verduidelikingsgaping ten opsigte van subjektiewe ervaring laat. In reaksie op hierdie ontologiese uitdaging stel die studie panpsigisme voor as ’n lewensvatbare alternatiewe grondslag vir AGI (KAI). Deur fenomenale eienskappe op ’n fundamentele vlak as ’n intrinsieke deel van materie te grond, los panpsigisme die probleem op van onbesielde ontstaan waardeur die fisikalisme gepla word. Daar word aangevoer dat hierdie metafisiese herformulering diepgaande implikasies het vir die AGI projek. Uiteindelik beweeg so ’n verskuiwing die doel van AGI verder weg, sonder om dit heeltemal te ontken. Die studie kom tot die slotsom dat panpsigisme ’n noodsaaklike ontologiese grondslag vir AGI is, waarin bewussyn verstaan word as ’n gebeurtenis van deelname en nie bloot as ’n fisiese proses nie.

## Isifinqo (Abstract: isiZulu)

Ngokomlando umkhakha wobuhlakani bokwenziwa i- (artificial intelligence) ubulokhu usekelwe emibonweni ye-computationalism mayelana nengqondo, okuyingxenye yemibono ye zazi zomzimba (physicalist) ephathelene nengqondo, lemibono ivame ukunganaki izindaba zephicamvelo(metaphysical) ezimayelana nokuqwashisa, ezivela ngenxa yokwamukela umbono wokuthi yonke into ingachazwa ngokomzimba kuphela. Nokho, njengoba ubuchwepheshe buqhubeka nokuthuthuka, umehluko phakathi kokulingisa ukuziphatha kanye nokokuqonda kwangempela kubonakala njengokumqoka kakhulu. Lolucwaningo luhlola izisekelo ze -ontoloji zobuhlakani bokwenziwa obujwayelekile (AGI), luveza ukuthi ukuthi imobono ebusayo ye-physicalist kanye ne- functionalist ayikwazi ukuchaza ngokugcwele isimo sokuqwashisa.Umqondo ka- David Chalmers' i-(Hard Problem of Consciousness) uyahlaziywa ukuze kuboniswe ukuthi izincazelo zengqondo ezisekelwe kuphela emisebenzini yayo ingqondo okanye esakhiweni sayo kazikwazi ukuvala igebe lokuchaza phakathi kwezinqubo ezibonakalayo zobuchopho kanye nesipiliyoni esizizwelayo somuntu.Ukuphendula kulenselelo ye-ontoloji, lolucwaningo luphakamisa i-panpsychism njengesisekelo esingaba enye indlela efanele yokwakhiwa kobuhlakani bokwenziwa okujwayelekile (AGI). Ngokubeka izimfanelo zesipiliyoni sokuqwashisa ezingeni eliyisisekelo lobukhona njengengxenye yemvelo yento uqobo,i- panpsychism ixazulula inkinga yokuvela okungachazeki okuhlupha umbono i -physicalism. Kuvezwa ukuthi lokhu kuhlelwa kabusha kwe-metaphysical kunemithelela ebanzi nebonakalayo kumklamo wobuhlakani bokwenziwa obujwayelekile (AGI). Ekugcineni, lolushintsho lwendlela yokucabanga lwenza umgomo wokufinyelela kubuhlakani bokwenziwa obujwayelekile (AGI) ubonakale ukude kakhulu kunalokho obekucatshangwa ekuqaleni, ngaphandle kokuphika ngokuphelele ukuthi ungafezeka. Lolucwaningo luphetha ngokuthi i- panpsychism iyisisekelo se ontology esibalulekile nesidingekayo ekwakheni umqondo wobuhlakani bokwenziwa obujwayelekile (AGI), lapho ukuqwashisa kungaqondwa khona njengesenzo sokubamba iqhaza kunokubhekwa njengenqubo nje yomzimba.

# Table of Contents

# Chapter 1: Introduction

## 1.1 Background

Consciousness has puzzled humans since the earliest attempts to understand the nature of being. The relationality between our conscious experience and the physical world forms the basis of the entire field of Philosophy of Mind. Our attitude towards this mystery affects not only our understanding of our own place in the world but also the moral standing of our co-inhabitants. René Descartes' formulation of the mind-body problem articulates these supposed dual aspects of reality.

> [On] the one hand, I have a clear and distinct idea of myself, insofar as I am merely a thinking thing, without extension, and, on the other hand, [I have] a distinct idea of body, insofar as it is merely an extended thing which does not think, it is certain that my mind is completely distinct from my body and can exist without it ([1641]1996, p. 5).

However, Descartes' dualism does more than separate the mental from the physical; it suggests an interaction between these disparate realms (Whitehead, [1927-28]1960, p. 10). In *Meditations on First Philosophy,* Descartes ([1641]1996, p. 79) uses the example of the mental experience of pain that accompanies a physical injury. The *problem of interaction* questions how these contrasting realms can influence one another.[1] If the mental and physical exist apart, how can thought lead to physical actions, or why does the physical world bring about mental sensations? Contemporary science points to a causally closed universe that does not allow for a "ghost in the machine" (Chalmers, 1996, p. 110).

The idea that our minds exist in some distinct realm from the physical world, therefore, seems at odds with science. Materialism, and more specifically physicalism, adheres to the laws of science and evades the problem of interaction by denying Descartes' proposed duality and, instead, explaining everything in the world as either physical or

[1] Elisabeth, Princess of Bohemia, is credited as the first person to articulate this problem in her correspondence with Descartes (Shapiro, 2021).

caused by a physical event (Chalmers, 1996, p. 42). Its intellectual roots trace back to early philosophical attempts to conceive the world as a unified system, rejecting the perceived contradictions of dualism (Lange, 1925, p. 47; Whitehead, [1927-1928]1960, p. 73; Stoljar, 2024). This unification necessarily leads to some form of identity claim (Smart, 2017).[2] *Token Physicalism* views "every particular thing in the world" as a physical entity (Chalmers, 1996, p. 147; Stoljar, 2024). In contrast, *Type Physicalism* claims that every property of the particular things in the world is reducible to some physical property (Chalmers, 1996, p. 147; Stoljar, 2024) Over the centuries since Descartes' mind-body separation, mechanistic physics and empirical psychology, driven by the progress of science, displaced the idea of a mental realm that is separated from the physical world (Stoljar, 2024). The prevailing *physicalist* theories see mental states as phenomena emerging from entirely material reactions within the brain, consistent with causal closure (Gibb, 2015, p. 17; Stoljar, 2024). This approach is commonly distinguished as a scientifically informed form of materialism since contemporary physics has revealed multiple entities and processes that are quite clearly physical, but do not qualify as "an inert, senseless substance, in which extension, figure, and motion do actually subsist" (Berkeley, [1710]2003; Stoljar, 2024). In what follows, I argue that this shift away from our dualistic conception of the mind effectively cuts subjective experience from our scientific understanding of the world.

Within contemporary philosophy of mind, the dominant approaches to understanding consciousness are still broadly physicalist (although that appears to be slowly changing) (Kuhn, 2024, Section 9). These varied approaches cover a spectrum of perspectives on how consciousness relates to the physical world. On the one end of the spectrum, *eliminativism* takes the maximalist position that consciousness does not exist or is some kind of illusion that is generated by the brain. Eliminativism is defined as "the thesis that our common- sense conception of psychological phenomena constitutes a radically false theory" (Churchland, 1981, p. 67). This "false theory" (i.e., the existence of consciousness) will eventually be completely disproven by neuroscience, according to eliminativists. Similarly, *illusionism*, while not rejecting consciousness outright, holds that certain properties of the mind "are an introspective

[2] It should be noted, as Kuhn (2024, Section 7) points out, "*any* theory of consciousness, to be complete and sufficient, must make an identity claim."

illusion" (Frankish, 2022, p. 322). On the other end of the spectrum, emergentism covers a range of approaches that take consciousness to be a novel property that is the result of the interaction of fundamental properties in a complex system (Chalmers, 1996, p. 49). This can be contrasted with reductionism by which everything can be explained by physics in principle, even if not in practise (Kuhn, 2024, Section 9). According to emergence in its weak form, emergent properties can be explained entirely by its fundamental properties in theory. Most contemporary philosophers are either implicitly or explicitly emergentists in some sense (Skrbina, 2020, p. 104). In its strong form (i.e., brute emergence), the emergent properties can never be explained by the fundamental properties even though they emerge because of them (Chalmers, 2016, p. 183).

Between these two poles on the physicalist spectrum lie many more approaches to explaining consciousness. *Functionalism*, for example, views mental phenomena as defined by the role that they fulfil within a system, rather than their substantive constitution (Block, 1978, p. 261; Putnam, 1992, p. 73). Because functionalism does not have such a strong token-identity claim as some of the other physicalist theories, it has become an important approach in the context of machine consciousness. Specifically, *Computational functionalism* "commits to the thesis that performing computations of a particular, natural and likely discoverable kind is both necessary and sufficient for consciousness" (Kuhn, 2024, Section 9). Chalmers (1995, pp. 200-202) challenges these physicalist approaches with *the hard problem of consciousness* by testing the assumption that subjective experiences can be reduced to or emerge from purely physical events in the brain. The hard problem lies in explaining the *subjective nature of conscious experience* itself (Chalmers, 1995, p. 201; 1996; 2003, p. 226). His argument forms the backbone of the one I intend to defend in this dissertation and will be discussed in more detail in the following chapter. Put simply, "easy problems" are concerned with aspects of consciousness investigable by cognitive science, involving understanding functions like perception, memory, and behaviour (Chalmers, 1995, p. 201). The hard problem of consciousness, in contrast, is concerned with the phenomenal element of consciousness, which lies outside the scope of those investigations (Chalmers, 1995, p. 201).

Reductively making sense of these so-called easy problems has led to remarkable progress in cognitive science and computer engineering. Alan Turing's (1950, p. 433) question, "Can machines think?" did precisely that by reframing "intelligence" as a computationally replicable function. As an academic discipline, the field of Artificial Intelligence (AI) commonly has its origins ascribed to a conference at Dartmouth College in 1956, where conveners aimed to expand on the concept of "learning machines" set in motion by Turing's (1950) paper *Computing Machinery and Intelligence* (Smuha, 2025, p. 1). The discipline slowly evolved out of advances in cybernetics and the formalisation of reasoning and learning processes into what it has become today (Marquis, Papini & Prade, 2020, p. 2; Russell, 2021, pp. 3-6). In the proposal for that conference, John McCarthy, Marvin Minsky, Nathaniel Rochester, and Claude Shannon ([1955]2006, p. 12) set out the aim for such an academic inquisition into AI as:

> An attempt to find how to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves.

There have been several different approaches to creating systems with such advanced capabilities, to varying and debatable levels of success.[3] AI research has focused on understanding "the principles underlying intelligent behavior" with the goal of recreating those principles in machines (Russell, 2021, p. 7).

The systems that now form part of our daily lives, designed to meet specific limited objectives such as Large Language Models (LLMs), voice assistants, chatbots, fraud detectors, facial recognition systems, etc. are examples of "narrow", "weak" or "specialised" AI (henceforth simply AI) (Dong, Hou, Zhang & Zhang, 2020, pp. 5-6; Searle, 1980, p. 417). These types of AI have been incredibly successful commercially because they are effective when given strict parameters, contributing to the general hype around AI (Dong et al., 2020, pp. 5-6; Mitchell, 2020, p. 87). "Broad", "strong", or "general" AI (Artificial General Intelligence or AGI), in contrast, artificial systems that "given the right programs can be literally said to *understand"* is yet to be achieved

[3] Testing for intelligence in artificial systems will be explored in more detail later in the study, but at present, it is enough to know that no system has been able to pass the benchmark Turing test to a broadly endorsed satisfactory level (Mitchell, 2020, p. 87).

(Dong et al., 2020, pp 5-6; Mitchell, 2020, p. 87; Searle, 1980, p. 417). In what follows, the focus will be on the second of these aspirational goals for AI, namely AGI, that takes "general intelligence" to be fundamentally different from specialised, task-serving competencies (i.e., narrow AI) (Goertzel, 2014, p. 1). To qualify as AGI, a system must be able to address novel challenges, navigate unexpected situations, transfer knowledge across domains, and, crucially, infer information from experience (2014, pp. 2-3). Specifically, the study will argue for the role of phenomenal consciousness in intelligent behaviour and the implications that it might have on achieving AGI with human capacities and behaviour. It is important to note that this is a connection that needs to be established, as intelligence and consciousness are not synonymous (Seth, 2025, p. 2). The distinction will be drawn out over the course of the dissertation, but broadly, *consciousness* is "something it is like to be" a conscious organism (Nagel, 1974, p. 436). *Intelligence* is a measure of capability or "ability to achieve goals" (Legg & Hutter, 2007, p. 23; Seth, 2025, p. 3).

In fact, Turing (1950, pp. 433-435) laid the foundation for engineering Artificial Intelligence by recognising the conceptual separation and shifting the objective from answering a vague philosophical question of "*can machines think?"* to a concerted effort to create a system that can be said to demonstrate intelligence in its behaviour when cross-examined alongside a human; one concerned with human-like output rather than process. The test circumvents the "mystery about consciousness", focusing on the human ability to recognise intelligence in others (1950, p. 445). Nonetheless, Turing's paper also perceives the role of human subjective experience in recognising intelligence in others (Dong et al., 2020, p. 3; Turing, 1950, p. 433). Around the same period, advances in formal logic, information theory, and cybernetics inspired the *computational theory of mind,* i.e., the idea that mental processes can be defined by the computational operations performed on symbolic representations (Newell & Simon, 1976, pp. 113-136; Pinker, 1997, pp. 83-86). As mentioned, computational functionalism underpins most AI research, from early symbolic systems to contemporary neural networks (Russell & Norvig, 2022, p. 50).

The contemporary methodological landscape of AI research, inspired by these questions about machine consciousness, is codified by Stuart Russell and Peter Norvig (2022, pp. 19-22) in *Artificial Intelligence: A Modern Approach* (AIMA). They

define the goals of AI research through four distinct lenses: "*acting humanly", "thinking humanly", "thinking rationally"*, and "*acting rationally"* (2022, pp. 19-22). This taxonomy (henceforth *the AIMA quadrant*) provides a practical framework for the functionalist project by distinguishing between behavioural output and rational modelling. However, the *explanatory gap* targeted by the hard problem of consciousness continues to trouble any reductive account of functionalism and highlights the importance of a sound philosophical ontology when considering AGI (Levine, 1983, p. 354). The Dartmouth conference proposal, mentioned earlier as a cornerstone of academic research into AI, makes one important stipulation:

> The study is to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it (McCarthy et al., [1955]2006, p. 13).

The problems faced by physicalism when confronting the "hard problem" call this fundamental assumption into serious question. If consciousness cannot fully be explained by physicalist theories, creating the capacity for it in machines to achieve true AGI may require a radically different framework. Admittedly, this argument assumes that AGI requires phenomenal capacities that still need to be defined in the coming chapters. Following John Searle (1980, p. 417), I aim to argue that it is precisely access to such subjective capacities that differentiates AGI from normal, narrow AI.

A possible alternative approach to the "hard problem", panpsychism, has seen a resurgence in contemporary philosophical discourse, spurred on, in part, by physicalism's inability to reductively explain the qualitative nature of experience (Chalmers, 1996, pp. 138, 274-282; Skrbina, 2009, p. 13; Goff, Seager, & Allen-Hermanson, 2022). Panpsychism is loosely understood as "the doctrine that everything has a mind" (Chalmers, 2015, p. 102). Contemporary panpsychists usually stop short of claiming that *everything* has the same conscious mind as we experience as humans; instead, the suggestion is that there must be a "mental" fundamental nature to some physical entities (Chalmers, 2015, pp. 102-103). These contemporary approaches can take various forms, distinguished by their respective descriptions of this fundamental nature. First, we can narrow it down by differentiating between two

main forms of panpsychism, namely *panexperientialism* and *pancognitivism* (Goff et al., 2022). Panexperientialism holds conscious experience to be fundamental (i.e., the what it is like experience), while pancognitivism claims thought is fundamental (i.e., "psychological attitudes towards specific propositions") (2022). Under a pancognitivist understanding, the argument for conscious AI would be as simple as: (1) if pancognitivism is true, everything has thoughts; (2) AI exists; therefore, (3) AI has thoughts.

That argument moves far too quickly, though, and we have no reason to believe rocks or electrons have a comparably complex mental life that we equate to thoughts. Instead, it is the panexperientialist interpretation that is concerned with the subjective phenomenal consciousness that is of interest to the research problem.

> According to the definition of consciousness that is dominant in contemporary analytic philosophy, something is conscious just in case there is something that it's like to be it; that is to say, if it has some kind of experience, no matter how basic. Humans have incredibly rich and complex experience, horses less so, mice less so again. Standardly the panexperientialist holds that this diminishing of the complexity of experience continues down through plants, and through to the basic constituents of reality, perhaps electrons and quarks. If the notion of "having experience" is flexible enough, then the view that an electron has experience—of some extremely basic kind—would seem to be coherent (Goff et al., 2022).

This idea that consciousness can be found at fundamental metaphysical levels has ancient roots in early Western thought (Skrbina, 2005, p. 21). Its resurgence is due in no small part to the fact that it allows us to give a coherent explanation for the role of consciousness (Skrbina, 2009, p. 13; Goff et al., 2022). As such, it will be argued that panpsychism provides the most parsimonious account of the mind. At this stage, a cursory introduction to panpsychism is enough to provide the background to the ontological questions from which the research problem arises. It is clear that such a shift in framework transforms the conceptual landscape of AGI research due to the physicalist assumptions underpinning contemporary approaches.

Currently, AI is experiencing a massive surge in interest (Dong et al., 2020, p. 1; Dežman, 2024, p. 743). Research around the world is focused on further developing

our AI technology. Companies like Google, OpenAI, and DeepSeek have made sensational advances, creating a multi-billion-dollar industry requiring governments and policymakers to develop formalised legislation to regulate rapidly advancing AI. Fears vary from extreme panic about an apocalypse that resembles a sci-fi movie to equally substantive, albeit more measured, concerns over the ethics of AI, job security, and the potential misuse of AI, to name but a few (Dong et al., 2020, p. 8; Dežman, 2024, p. 743). There is also a sense of urgency among researchers, governments, and businesses to be among the earliest adopters, recognising the untapped potential for AI as a productivity tool that could revolutionise human development and social progress (Dong et al., 2020, p. 8). There are issues concerning privacy and data protection, fairness and bias, and the general relevance of technology for the human condition that all need to be considered (Müller, 2026).

These are all fascinating and pressing questions. This dissertation, however, is focused specifically on the ontological divide illustrated by Chalmers' challenge. Since we have no other window through which to observe human intellect other than our own experience, we must rely on introspection to recognise phenomenal consciousness. Critics often highlight that human memory and introspection are faulty mechanisms, raising the question of how we can reliably base an ontological framework on such a fallible access point (Dennett, 1979, p. 95). However, while our introspective lens may be prone to cognitive bias or misidentification regarding the *causes* of our mental states, it consistently registers the baseline reality that subjective experience exists (Chalmers, 1996, pp. 189-190). It is not easy to conceive how a physicalist ontology that cannot account for this baseline phenomenon can bring about a complete account of human intelligence if it is accepted that subjective phenomenal elements have a part to play. If physicalism or functionalism is proven false, an account of this subjective experience appears to be a prerequisite to recreating human capacities in an artificial system.

Understanding this context is crucial for what follows. The dissertation does not advocate for any particular technological agenda; its task is conceptual clarification. The argument proceeds from the observation that existing physicalist and functionalist frameworks cannot account for phenomenal consciousness to the claim that

panpsychism provides a more coherent ontological grounding, all within the broader context of AI research.

## 1.2 Research question

Throughout this study, the question that will be posed as the lodestar is:

> *Given the difficulty in explaining the hard problem of consciousness from a physicalist perspective, does panpsychism provide a viable foundation for achieving AGI?*

We have set a goal of creating artificial systems with human-like reasoning and behavioural capacities. To achieve this and to determine how to regulate and assign moral status to such potential systems, we require a robust account of the mind. The research question challenges the viability of a physicalist ontology to provide such an account of the mind, including phenomenal experiences. In its stead, panpsychism is proposed as an alternative that places consciousness at a fundamental level. The core argument will be that there are phenomenal aspects of consciousness that have inherent value in terms of the capacities we associate with general intelligence. If physicalism cannot explain these phenomenal qualities, AGI requires an ontological foundation that can.

## 1.3 Aims and objectives

This dissertation aims to critically explore the potential of panpsychism as an ontological framework for hypothesised AGI. It will be argued that panpsychism presents a more convincing explanation of the mind than prevailing physicalist theories. Additionally, it will be argued that subjective experience plays a crucial role in what we expect from AGI. By contrasting physicalism and panpsychism as ontological frameworks for understanding the mind, the study will consider the possibility of AGI within the contrasting approaches and evaluate whether existing theories of consciousness are theoretically capable of providing a blueprint of the mind to recreate in artificial systems.

To achieve this, the following objectives have been derived from the research question:

1. To show that the hard problem of consciousness presents a fundamental challenge to physicalism.
2. To establish the role of phenomenal consciousness in the capacities we hope to see in AGI.
3. To critically assess alternative metaphysical frameworks for consciousness, focusing on panpsychism.
4. To argue that panpsychism transforms the conceptual landscape of AGI research.
5. To consider the philosophical, ethical, and epistemic implications of panpsychism for future AI development and cognitive research.

## 1.4 Rationale

The research question is inspired by the continued pursuit of AGI. Steady progress in cognitive science and ongoing technological advances give the problem of consciousness new urgency. As increasingly complex AI systems demonstrate advanced linguistic, perceptual, and reasoning abilities, the distinction between behaviour and experience becomes conceptually strained. The study aims to clarify how the hard problem of consciousness complicates claims about AGI and why panpsychism may provide a more coherent understanding of both.

Contemporary AI is based on a computationalist understanding of the mind that identifies cognition with information processing. Building on the functionalist legacy of mid-twentieth-century philosophy, this approach defines mental states according to their place in the causal chain, rather than their substrate or structure (Pinker, 1997, pp. 83-86; Putnam, 1997, pp. 31-44). This framework has allowed for remarkable empirical success, culminating in the architecture that underpins the systems we know today. Yet, such systems, despite their impressive functional abilities, are not considered to qualify as AGI. They seem to lack the “something it is like” to be them,

as Thomas Nagel (1974, pp. 435–456) famously phrased the problem of subjective experience.

This dissertation recognises the disjunction between the functional and the phenomenal as more than a scientific mystery to be solved, but an ontological problem that forces us to revisit our most fundamental assumptions. The difficulty arises from the implicit physicalist premises that dominate philosophy of mind and AI research. The significance of this limitation is philosophical rather than technical. It shows that the explanatory gap between brain and experience (or between computation and consciousness) does not arise from incomplete science but from a metaphysical framework that excludes subjectivity from the start. The hard problem thus reveals an ontological inadequacy in the physicalist conception of nature. To think coherently about consciousness, whether human or artificial, this underlying assumption must be reconsidered.

> Unlocking AI's barrier of meaning will likely require a step backward for the field, away from ever bigger networks and more massive data collections, and back to the field's original roots as an interdisciplinary science studying the most challenging of scientific problems: The nature of intelligence (Mitchell, 2019, p. 51).

Addressing this "barrier" may require a shift away from the prevailing approaches that seem to focus on scaling processing power and resources, returning instead to the interdisciplinary roots of AI research and the fundamental question of what intelligence is. As explained in the background section, this study is concerned with Artificial Intelligence in the broad sense. Accordingly, it seeks an acceptable account of the subjective nature of experience necessary for reverse-engineering machines capable of approximating rich and complex human mentality.

Panpsychism offers one way of doing so. Rather than attempting to explain consciousness as the emergent result of complex processes in some physical systems, panpsychism holds that experience, in some primitive form, is a "fundamental and ubiquitous" feature of reality (Goff et al., 2022; Skrbina, 2009, pp. 1-29). Since our only direct acquaintance with intrinsic properties comes through consciousness itself, it is plausible to argue that intrinsic nature is identified with phenomenal experience. From

this perspective, consciousness is not an anomaly that emerges within an otherwise insentient universe, but a continuous aspect of the natural order (Goff et al., 2022). This reframing carries clear implications for AI, as I will show.

Mitchell's (2019, p. 51) call for a return to foundational inquiry aligns with the rationale of this dissertation. The hard problem of consciousness resists being solved through reduction alone. It requires conceptual clarity about what kind of thing consciousness is and how it fits within the fabric of being. The current dominance of data-driven methodologies has obscured this metaphysical dimension, treating consciousness as either irrelevant or reducible (Churchland, 1981, p. 67; Dennett, 1991, pp. 285-314). However, Searle (2008, pp. 73-74) reminds us that consciousness is characterised by a subjective experiential aspect that is not compatible with the objective, descriptive framework of science. Therefore, it is necessary to reevaluate the ontology that informs our scientific models and metaphysical theories to understand the nature of the mind.

The rationale for this dissertation is to precisely pursue this ontological clarification. Its purpose is to articulate the conditions under which claims about consciousness – human or artificial – can be philosophically coherent, and not to attempt to empirically verify or develop a novel theory of consciousness for AI. To do this, the hard problem of consciousness is used to demonstrate a shortcoming of physicalist frameworks and to show why panpsychism offers a favourable metaphysical framework that is conceptually wide enough to encompass the phenomenal aspects of consciousness. The goal is to contribute to the discussion in the philosophy of mind and AI by explaining why our understanding of intelligence needs to be philosophically grounded.

The question of consciousness – whether attributed correctly or not – in AI may not be merely speculative. As large-scale machine learning systems increasingly improve in mediating tasks previously thought to require human intelligence, the conceptual boundaries between human and artificial intelligence become ever more blurred. Words like *understanding*, *reasoning,* and *creativity* are already routinely used when describing purely algorithmic machine operations. Without a clear account of consciousness, such language risks equivocation. The philosophical responsibility, therefore, is to articulate the conditions that make the attribution of consciousness

meaningful, not only metaphorical. In this context, the dissertation partakes in the growing dialogue between philosophy, cognitive science, AI engineering, and psychology by treating consciousness as a fundamental metaphysical question with practical implications. It argues that AGI, if it is ever to be realised in a meaningful sense, must be understood not only as an engineering feat but as an ontological event that would require reconceiving the relationship between mind and matter. Panpsychism provides the framework for such reconception.

Ultimately, the rationale for this study is grounded in the conviction that the nature of consciousness cannot be approached solely through the language of mechanism. The physicalist paradigm has advanced our descriptive and predictive understanding of intelligence, but it has done so by abstracting away the very phenomenon it seeks to explain. To move forward, we must first step back—away from technical accumulation and towards philosophical clarity. In this sense, the present research joins the broader effort to reintegrate philosophy into the study of mind and intelligence. Only by understanding what consciousness *is* can we responsibly inquire into whether it might one day exist in artificial form.

## 1.5 Methodology

As this study primarily engages with the ontological question of the mind, it is placed firmly within the field of metaphysics. However, we find ourselves at the precipice of a potential paradigm shift, where the lines between philosophy and its sister fields are once again fading out of existence.

> Look at what subtle philosophical distinctions about thought processes have suddenly become relevant to a discussion about cutting edge technology! We could justly say that we live in an age in which Socrates has come to be in dialogue with Hephaistos [the Greek god of craft]. What would have seemed like pure theoretical philosophy has become continuous with the bleeding edge of engineering (Koralus, 2024, p. 4).

Since the research question spans several different academic domains (i.e., neuroscience, psychology, engineering, biology, and physics), the study will inevitably veer into these different lanes. In keeping with the objectives set out above, the study

will attempt to resolve ambiguities across the disciplines and interpret the information to address the research question. To achieve this, the study employs a conceptual and comparative philosophical methodology situated within the field of metaphysics, with a focus on philosophy of mind and artificial intelligence.

The research adopts a qualitative, interpretive, and analytic approach. As a philosophical investigation, this study involves a secondary examination of the literature and an assessment and synthesis of its arguments. Because the inquiry is theoretical and interpretive, no human participants or experimental data are involved. Instead, the research synthesises existing philosophical and technical literature to evaluate whether the metaphysical commitments of AI research require revision.

In carrying out the research, reference is made to academic books and journal articles, lectures, conference papers, electronic resources, and magazine articles, and engaging in advisory discussions.

## 1.6 Ethical Considerations

This project does not contravene any regulations in the UNISA Disciplinary Code for Students, as all necessary ethical considerations have been taken to ensure that the project has been conducted fairly and honestly per the University's Policy on Copyright Infringement and Plagiarism. As this is a purely theoretical study following a qualitative philosophical methodology, no human or non-human animal subjects have been involved or negatively impacted.

In compliance with the Unisa Policy on Research Ethics and the Standard Operating Procedure on Research Ethics Risk Assessment, ethical clearance was obtained from The College of Human Sciences_CREC.

NHREC Registration #: (Rec-240816-052)
Ref #: 6114

## 1.7 Overview and outline

To address the research question and achieve the overarching objectives identified above, the dissertation begins in Chapter 2 with a detailed examination of the philosophical landscape that gives rise to the hard problem of consciousness. As such, the chapter draws heavily on David Chalmers' (1996) *The Conscious Mind* to describe the impasse faced by physicalist and functionalist theories when challenged to explain the phenomenal aspect of consciousness as characterised in human experience. It analyses fundamental commitments of physicalism, functionalism, and computationalism, and shows why these frameworks fail to explain consciousness in its phenomenal sense.

By reviewing key arguments such as Nagel's (1974, pp. 435-449) "what-it-is-like" challenge and Frank Jackson's (1982, p. 130) *Knowledge Argument*, the chapter demonstrates that a purely structural/functional description of the mind leaves an explanatory gap regarding subjective experience. Using the hard problem of consciousness, this gap is shown to be a result of an ontological inadequacy in our worldview, rather than merely a technical challenge to be solved by further reduction. To anchor this insight in the context of AI, I consider the conceivability of conscious AI through a thought experiment.

Recognising the conceptual distinction between intelligence and consciousness, Chapter 3 begins by establishing the role of phenomenal consciousness in the capacities, as yet reserved only for humans, that represent the target of AGI. Following Jackson (1982, p. 130) and Searle (1980, pp. 417-420), the inherent significance of consciousness is argued to be the missing ingredient in current approaches to AI. Bridging these philosophical insights with the domain of AI engineering, the third chapter examines how prevailing AI paradigms implicitly depend on physicalist and functionalist assumptions. An analytical review of existing theories of consciousness reveals that any framework that excludes the intrinsic nature of consciousness will fall short of providing a blueprint for instantiating a mind, restricting us to narrow AI.

Responding to the perceived failure of reductive frameworks, Chapter 4 evaluates alternative metaphysical positions better suited to incorporate phenomenal consciousness in their scope. First, *Naturalistic Dualism* is assessed, focusing on its

inherent struggle with the threat of epiphenomenalism. Finding responses to the interaction problem unsatisfactory, the enquiry transitions toward *Russellian Monism*. Bertrand Russell (1959, p. 18) and Arthur Eddington's (1928, pp. 259-260) respective commentaries on the limitations of physical theory are employed to draw out the distinctions between intrinsic and extrinsic properties. Thereby, the concept of underlying, non-physical properties – i.e., "quiddities" – is introduced, which sets the foundation for the focal example variant of panpsychism that will be investigated.

Building on said foundation, the argument from intrinsic nature and the anti-emergence argument will be considered as the two most compelling motivations for panpsychism, along with some more speculative motivations concerning compatibility with historically marginalised and forgotten schools of thought. In keeping with the critical approach to the research question, the *combination problem* is then identified and assessed as the most puzzling challenge for panpsychism. In doing so, panpsychism is established as the most promising ontological alternative to physicalism for explaining consciousness. I make the argument that if consciousness is fundamental, then AGI cannot be understood merely as an emergent property of computational complexity but as the potential organisation of matter already possessing intrinsic experiential properties. This re-conceptualisation affects our understanding of intelligence and what would be required to artificially instantiate it in a machine.

Finally, the findings are synthesised in Chapter 5 to offer concluding remarks on the philosophical, ethical, and epistemic implications of the proposed metaphysical shift. I reiterate the impassable barrier physicalism faces when accounting for phenomenal experience and the fatal implications thereof for the AGI project. In physicalism's stead, panpsychism is proposed as a way forward, albeit with the caveat that it complicates the project even further.

# Chapter 2: The hard problem of consciousness for AI

## 2.1 Introduction

The research question outlined in the previous chapter can be broken down into two distinct elements that will need to be explained before the argument begins in earnest. First, the claimed difficulties faced by physicalism when faced with the hard problem of consciousness must be described. Once this has been achieved, attention will shift to the second part of the research question, in the following chapter, regarding panpsychism.

This chapter engages heavily with Chalmers' (1996) argument against materialism, which has become known as *the hard problem of consciousness*. To begin, consciousness needs to be defined and placed within the mental realm as described by the Cartesian duality introduced in the background section. Following Chalmers, Nagel's *What is it like to be a bat?* paper (1974, pp. 435-449) is referenced to further differentiate between the subjective and objective nature of the mind. It is these subjective experiential features of consciousness that lay the foundation for Chalmers' argument for the inadequacy of physicalism.

Jackson's *Knowledge Argument* (1982, p. 130) elucidates the type of consciousness Chalmers is concerned with. To understand where this phenomenal sense of consciousness fits in the natural order, the concept of *supervenience* as defined by Chalmers (1996, p. 32) needs to be explained. Here, some groundwork in the philosophy of language and modal logic from Saul Kripke's *Naming and Necessity* lectures (1980) will help to understand Chalmers' *Zombie Argument* that challenges the idea that physical facts supervene mental facts (1996, p. 94).

In the next section, some physicalist objections in the form of *Illusionism*, *Eliminative Materialism*, and *Naturalistic Physicalism* will be considered, along with Chalmers' responses (Churchland, 1981, p. 67; Dennett, 1991, pp. 285-314; Papineau, 1993[2013]). I will explain how these respective approaches to physicalism attempt to

explain the subjective aspects of consciousness before showing that they inevitably run into the same problems.

Since the research question is specifically interested in the consequences of the hard problem of consciousness in the context of the development of AGI, as outlined in the background section, section 2.4 establishes the conceivability of conscious AI. Returning to the Turing test, Searle's (1980, pp. 417-420) thought experiment is used to show the distinction between functional output and understanding. With this in mind, the concept of AI zombies is introduced to illuminate the problems AGI holds for physicalism in terms of the subjective nature of consciousness. Through this, I intend to argue for the place and relevance of phenomenal experience when describing consciousness and to show why AGI requires a complete account of the mind.

## 2.2 The hard problem of consciousness

To understand the hard problem of consciousness, it makes sense to begin by defining what is meant by "consciousness". Chalmers (1996, p. 3) characterises consciousness as "at once the most familiar thing in the world and the most mysterious". Ned Block (1995, p. 227) calls consciousness a "mongrel concept".[4] Both refer to the fact that consciousness is commonly used as a catch-all term for a wide array of features of the mind and human experience, and that such a conflation of disparate concepts and phenomena leads to much of the confusion on the topic.

Consider what comes to mind when hearing the term "conscious". It could refer to "cognitive capacity" (i.e. being able to reflect on or report one's own mental state) (Chalmers, 1996, p. 6). It could be used to indicate a state of "awakeness" (1996, p. 6). It can also be used in the sense of being able to focus one's attention, or the wilful control of one's behaviour (1996, p. 6). Being "conscious of something" could simply

[4] Block differentiates what he means by "mongrel concept" from "cluster concept" (Block & Manson, 1998, p. 36). Cluster concepts group varied, yet related, conceptual elements together. Mongrel concepts differ in that they are characterised by grouping conceptual elements of completely different sorts with "[1] no scientific unity" and which "[2] promotes conflation" (1998, p. 37). Religion is a cluster concept because it encapsulates different elements. In comparison, speed is a mongrel concept because it lumps together two distinct conceptual elements, namely "average and instantaneous velocity" (1998, p. 36).

mean “to know about something” (1996, p. 6). However, in the context of this study, consciousness refers specifically to the way in which we humans experience the world. Consciousness, in this sense, has a distinctly phenomenal character. Nagel (1974, p. 519) describes this as the “something it is like” quality of consciousness.

In “*What is it like to be a bat?”,* Nagel (1974, pp. 435-449) challenges us to imagine a bat’s subjective experience of being and shows that no reductive account, regardless of complexity, can capture the essence of the experience of “bat-ness” that accompanies the physical facts about a bat.[5] In doing so, Nagel exposes the shortcomings of a reductionist account of the mind. To draw out the distinction between what he calls the “*pour-soi”* (for itself) and the *“en-soi”* (in itself), Nagel (1974, pp. 437-438) uses the example of a bat’s presumed inner experience. Being a mammal, we can assume that bats have some semblance of consciousness, at least to the level of something like a dog or a rat (1974, p. 438). However, unlike examples that share similar neural activity and sensory apparatus with humans, Nagel (1974, p. 438) explains that bats present a particularly vivid illustration of the problem he is considering because they experience the physical world through sonar. By observing the inner workings of bats, we learn objective facts (i.e., for itself) about the physical processes involved in such a sonar experience. Reductive reasoning and empirical investigation can reveal a range of objective truths about what happens in a bat’s mind; bats use their rapid, subtly modulated shrieks to generate high-frequency waves that echo off the external physical world, which they can detect and correlate to specified outgoing impulses to form a precise informational picture of the world (Nagel, 1974, p. 438). Even though we can understand these processes involved in a bat’s experience, it brings us no closer to subjectively (i.e., in itself) understanding *what it is like* for the bat to experience the world (1974, p. 438).

Here is a thought experiment by way of an example. From where I am writing this, sitting at my desk in the corner of my office, with music playing in the background, I can observe my conscious experience introspectively. The first thing I notice is that I am aware of my sensory experience: the light from the screen of my laptop, the sound

[5] Joseph Levine (1983, p. 355) refers to this as the “felt contingency” about the physical facts that accompany experience.

coming through the speakers. I differentiate between shapes and colours and form an image of the world in front of me, infused with sounds and smells, and tactile feel. Therefore, I can be said to be conscious of the world. This in itself suggests that I am alive and awake, thus conscious in the world. On the most superficial level, all of these conscious experiences are easily explainable by bodily processes. Looking a little bit deeper, I can be conscious of my body and certain sensations, like a dull pain from where a bee stung me earlier, or a sense of energy because I have just consumed a sugary drink. I am in a particular emotional state. Intrigued by the question at hand, stressed about completing this draft. The words on the screen have meaning to me because of the inferences I can make based on my understanding, learning, and memory. The vibrations coming through the speakers on my laptop, similarly, become tones that form the song I am hearing. This, in turn, has the further effect of inspiring imagination or recalling memories and emotions. In my particular case, there is an added noise which is not caused by these external vibrations, but by a damaged auditory system caused by an accident I was involved in as a child, which left me with tinnitus ever since. This is a constant, high-pitched buzzing noise in my right ear. Sometimes I am conscious of it, and other times I seem not to notice it. However, as soon as I focus my attention on it, the buzzing can almost drown out the external noise. On a meta-level, I also seem to be conscious that I am a unique entity within an external world who is having this experience on a private, internal level. Finally, the fact that I am considering my own consciousness in this way suggests that introspection, reportability, and voluntary control are features of consciousness. This is nowhere near a complete picture of my conscious experience at this moment, but it should be enough to illustrate the point.

Questions about these aspects of consciousness that are concerned with the functional or explanatory role of mental states all fall under what Chalmers (1995, p. 202; 2007, p. 225) calls “the easy problems of consciousness”. Consider some more examples of these functional elements of consciousness, like wakefulness, voluntary control, self-consciousness, reportability, and awareness (1996, pp. 26-31). Being awake can be analysed functionally as “the ability to process information about the world and deal with it in a rational fashion” (1996, p. 26). Voluntary control is exemplified by deliberate action, which in itself is a structural response to a thought (as in neurobiological activity) (1996, p. 27). Similarly, self-consciousness requires

access to a "self-model" (1996, p. 27). Self-consciousness can be reductively understood as structural possession of such a reflective inner representation (1996, p. 27). Reportability presupposes some structural capacity for language (1996, p. 26).

"Awareness" broadly encapsulates most, if not all, of the functional aspects of consciousness described above (Chalmers, 1996, p. 29). "Briefly put, we can think of awareness as direct availability for global control" (1995, p. 212). In other words, a system is aware when it has access to information that can be used to control the system's behaviour (1996, p. 28). Block (1995, pp. 230-235) refers to the aspects discussed above as "Access-consciousness" or "A-consciousness", which he distinguishes from "P-consciousness" (Phenomenal consciousness). For the current discussion, Chalmers' "awareness" can be taken as synonymous with Block's "A-consciousness".

Block defines A-consciousness as a cluster concept representing a state which is "poised for use as a premise in reasoning" (i.e., available for decision making), "poised for rational control of action", and "poised for rational control of speech" (1995, p. 231). Of these elements, he describes the third – reportability, as the one which carries the least weight in the cluster, yet "is often the best practical guide to A-consciousness" (1995, p. 231).[6] Put simply, the content of P-consciousness is phenomenal, whereas the content of A-consciousness is representational (1995, p. 232).

For Chalmers (1996, pp. 17-22, 29), this distinction between P-consciousness and A-consciousness tracks the distinction he draws between the phenomenal aspects of consciousness and the reducible aspects of consciousness. When considering consciousness, it is easy to focus only on the latter. Under the functional conception of the mind, a state is "mental" if it performs a causal or explanatory role in behaviour (1996, p. 11). This includes all neural processes involved and accounts for many of the phenomena we associate with consciousness.

[6] Block (1995, p. 231) regards reportability as "not necessary" and "not independent of the others" because he argues that non-linguistic animals also possess A-conscious states.

> On the phenomenal concept, mind is characterized by the way it *feels;* on the psychological [i.e., functional] concept, mind is characterized by what it *does* (Chalmers, 1996, p. 11).

According to Chalmers (1996, p. 25), "the psychological mind-body problem has been dissolved."[7] This is because cognitive science provides explanatory models for these phenomena (1995, p. 200; 2007, p. 225). Even emotions can be explained in terms of structural coherence (1995, p. 213). Though there is a clear phenomenal aspect attached to emotions in the way that we think about the feeling of emotions more immediately than the physical processes involved, there is a definite objective element in the chain of neurobiological events that accompany them. Truths about the information processing (i.e., structural, functional, objective) elements involved when experiencing love or pain or joy are empirically available, but that does not paint the full picture.

Looking at the example snapshot of my experience, we can differentiate between the phenomenal and the functional (i.e., the "psychological" or computational) aspects of mind (Chalmers, 1996, pp. 26-28). The divide is not always self-evident. Upon first consideration, the intuition might be that everything I have described can be encapsulated and described by the awareness capacity mentioned above. This is because the phenomenal and physical aspects are usually co-occurring as "conscious experience does not occur in a vacuum" (1996, p. 22). The glow of the screen, the sound of the music, the aching sensation in my ear, and the persistent tinnitus all have a clearly phenomenal aspect in that there is a certain subjective quality to the experience. All of the information processing and causation happening in that moment "does not happen in the dark" (Chalmers, 1996, p. 4). We know that when I am listening to music or looking at a screen, I have internal systems reacting to external stimuli. The process from pixels projecting into my retina, or the sound waves reaching my eardrum, can be scientifically analysed to provide a comprehensive account of every process involved. This goes no further towards explaining why my phenomenal

[7] Chalmers (1996, p. 11) uses the term "psychological" rather than "functional". To avoid confusion with the well-established science of psychology, I will refer to the functional concept of mind when discussing Chalmers' psychological concept of mind. By Chalmers' (1996, p. 11) definition, "[a] state is mental in this sense if it plays the right sort of causal role in the production of behavior, or at least plays an appropriate role in the explanation of behavior". This change does not affect the argument; it simply helps to avoid confusion.

experience should feel like they do – “Why should it feel like anything at all?” (1996, p. 7).

The example of pain has been used already. It is perfectly explainable what has happened during and after the event of the bee stinging me to bring about the experience of pain. However, I was not describing those processes when observing the ache in my ear; I was pointing to the unpleasant phenomenal quality of being in pain. Clearly, there are two aspects of pain that both seem equally central to the concept. The functional aspect of the experience is the physical brain-state that is produced when the organism is damaged and usually causes some form of aversion reaction. Nothing about that aspect explains the co-occurrence of the phenomenal aspect; the best we can do is notice that they both seem central to our conception of pain.

> One can tie oneself into all kinds of knots by worrying about whether the phenomenal quality or the functional role is more essential to pain. For instance, would a hypothetical system in which all the functional criteria were satisfied but in which the conscious [i.e. phenomenal] experience were not present be truly in pain? One might be tempted to say no, but what of the fact that we speak of pains that last for a day, even though there are times when they are not conscious? There is little point trying to legislate matters one way or the other. Nothing important rests on the semantic decision as to whether some phenomenal quality is really essential for something to count as pain. Instead, we can recognize the different components associated with a concept and explicitly distinguish them, speaking for example of ‘phenomenal’ pain and ‘psychological’ pain. Our everyday concept of pain presumably combines the two in some subtle weighted combination, but for philosophical discussion things are clearer if we keep them separate (Chalmers, 1996, p. 17)

It might be easy to conflate awareness with the phenomenal consciousness this study is focused on because they seem to accompany one another in nearly any conceivable example (Block, 1995, pp. 230-245; Chalmers, 1996, pp. 22-24; 28). To draw out the separation, Block (1995, pp. 243-244) refers to examples of medical procedures where patients experience the feeling of pain (i.e., P-consciousness) without awareness in the functional sense (i.e., A-consciousness) of the information in a way that can inform

a reaction.[8] Chalmers (1995, p. 202) and Block (1995, p. 236) agree that it is a misunderstanding of this crucial nuance between awareness and phenomenal consciousness that leads to much of the confusion on the topic.

Clearly, some of the other events in the example snapshot of my conscious experience, such as the thoughts and memories when reading the words, my emotions, and the sense of self, also have distinct phenomenal qualities. They feel like something. There are functional processes happening when they occur, but that is a philosophically distinct subject from the one this dissertation is interested in. Other aspects of my consciousness, such as wakefulness, voluntary control, self-consciousness, reportability, and awareness, have all already been considered from a functional perspective above. Yet there is a certain phenomenal state that accompanies these functional processes.

The tinnitus presents an especially interesting case. At first glance, it seems to be completely phenomenal in that there is no sound wave causing it. However, scientists have been able to make some progress researching tinnitus, even if it remains puzzling (Baguley, McFerran & Hall, 2013, p. 1605). Research shows that even after the auditory nerve has been severed in test subjects, the sound persists (2013, p. 1601). The precise neural correlates of the phenomenal experience of tinnitus remain difficult to confirm (2013, p. 1601).

The point is not to try to argue that the phenomenal aspect of tinnitus is any more or less relevant than the functional role behind it. Rather, it is to show that the phenomenal experience is not as straightforwardly connected to physical function/process as we might assume. Chalmers (1996, p. 32) demonstrates this through investigating the concept of *supervenience* that underlies a physicalist account of the mind.[9] The core idea of the concept, as employed by Chalmers (1996,

[8] More formally, Block (2007, pp. 487-489) has differentiated between A-consciousness and P-consciousness using the *Phenomenal Overflow Argument*. Essentially, he makes the case that we humans may have some P-conscious states that are inaccessible to A-consciousness, proving that they are distinct. This tangent lies outside the scope of this paper and Chalmers' argument does not depend on it.

[9] The term "supervenience", as a philosophical concept, is generally credited as being introduced by R.M. Hare, though it appears to have been in use in its current form by his contemporaries already (Kim, 1990, pp. 3-5).

p. 33), is that “B-properties” (i.e., mental properties) are taken to supervene on “A-properties” (i.e., physical properties) if it is inconceivable to have a situation where “A-properties” are identical while “B-properties” differ. In Chalmers’ argument, B-properties correspond to the phenomenal experience of a mental state, and A-properties include the physical processes that occur. The supervenience claim is that any two agents that share the same physical processes would have the same phenomenal experience (Kim, 1990, p. 7). Here, Chalmers makes a key point to fit his modal argument. To do this, he draws on Kripke’s (1980, pp. 33-39) groundwork on the necessary *a posteriori*.

According to Kripke (1980, pp. 137-139), some necessary truths are known *a posteriori*. To explain properties that are essential to its referent, he uses the notion of “rigid designation” (1980, pp. 10-11). Essential properties are characterised as those that belong to an object that could not exist without having that property. If something is rigidly designated, it points to the same entity in any possible instantiation in any possible world (Kripke, 1980, p. 48). Kripke (1980, pp. 109-110) uses the example of the planet Venus being designated by the names *Hesperus* and *Phosphorus* in Greek mythology. Because both terms, Hesperus and Phosphorus, rigidly designate the planet Venus, the empirical discovery that Hesperus = Phosphorus, if true, must be necessary (i.e., it must be universally true). We might learn this necessary truth (Hesperus = Phosphorus) only *a posteriori* (after empirically discovering that both names designate the planet Venus).

Chalmers (1996, p. 41) uses this insight to inform his mind-body argument, stating that “B-properties” can be said to supervene on “A-properties” whenever “A-facts” entail “B-facts”. Recalling the physicalist thesis defined in the background section that considers mental states to be identical to physical states, a logical argument is required to explain why mental phenomena must supervene on the physical in a strong global sense. To give a reductive account of a phenomenon, the phenomenon itself must be “globally logically supervenient” on the lower-level properties being used to explain it (Chalmers, 1996, p. 48). In Kripke’s (1980, p. 98) terms, physicalists consider the phenomenal experience of pain and the accompanying material state of the brain (such as the stimulation of nerve C-fibres) to rigidly designate the same thing. Put simply, if physicalism is true, two systems that are physically and functionally identical

must possess identical conscious experience because consciousness is taken to be supervenient on the physical.

Jackson (1982, p. 130) challenges this assumption in the *"Knowledge Argument".* In the thought experiment Jackson uses, "Mary" is a scientist familiar with a complete neurophysiological account of vision (i.e., a perfect understanding of every physical process involved with perceiving colour). However, Mary exists in a world where she has experienced only black and white. Suppose that Mary sees a purple flower for the first time; does she gain any added information about the world? According to Jackson (1982, p. 130), it seems "just obvious" that Mary does indeed "learn something" new about the world through her experience of the phenomenal aspect of perceiving colour. If physicalism holds, then a complete physical description should leave no fact unaccounted for, and there would be no new knowledge for Mary to gain. At the time, this was enough for him to conclude that "[p]hysicalism is false" (1982, p. 130).[10] For this study, however, the knowledge argument serves to illustrate the epistemic value of subjective phenomenal experience. The *purpleness* of purple, not in terms of light and wavelengths or hue and chroma as a value on the electromagnetic spectrum. Or the feeling that accompanies listening to an evocative song, not the vibrations absorbed by sensory cells in the ear and converted to signals processed through the auditory nerve and interpreted by the brain.

Block (1978, pp. 304-310) argues that the existence of these intrinsic mental features of experience (i.e., P-consciousness) challenges the functionalist account of the mind. To demonstrate that these P-conscious states are not identical to the A-conscious (i.e., functional) states that accompany them, Block examines cases in which such P-conscious states are either *inverted* or *absent*. If we can conceive of a system with an identical functional organisation to that of our own, without P-consciousness, "then consciousness cannot arise from functional organisation" (Chalmers, 1996, p. 250).[11]

[10] Jackson (2003, pp. 261-263) later revises his position to embrace physicalism. To explain why we might erroneously conclude that Mary gains new knowledge upon her first experience of colour, he suggests that Mary gains a new ability to *represent* knowledge, not new *knowledge*. However, the epistemic gap at the centre of this study remains unaccounted for (Chalmers, 2003, pp. 106-108).

[11] This is commonly referred to as the "Absent Qualia Argument" (Block, 1978, p. 281). I avoid the term "qualia" because of all the conflating literature on the concept (see Tye (2025)). In the context of this study, *qualia* can be taken as synonymous with P-consciousness or the subjective element of consciousness.

Block (1978, p. 279) uses the example of recreating the functional output of a human mind using volunteers (the population of China in Block's original example) to mimic the output of neurons in the brain. While such a system does not seem physically impossible, it is not self-evident that consciousness will emerge from its formation (Chalmers, 1996, p. 97). The *inverted spectrum* argument is based on the hypothesis that inverted P-conscious states may accompany identical A-conscious states in two different agents.[12] In other words, where two people might point towards the same object and both call it red, one might subjectively be experiencing red while the other experiences green. The conceivability of such a scenario challenges the functionalist account of the mind because "as long as some positive fact about experience in our world does not hold in a physically identical world, then consciousness does not logically supervene [on the physical]" (Chalmers, 1996, p. 99).[13]

These cracks in the physicalist claim are fully exposed by Chalmers' (1996, p. 94) *Zombie Argument*. The thought experiment asks us to imagine a reality in which all the physical facts, including the molecular and "psychological"[14] makeup and behaviour of its inhabitants are identical to our own reality. The only difference is that agents in this world (i.e., "zombies") lack subjective conscious experience (1996, pp. 94-95). Chalmers contends that the conceivability of such a reality implies that consciousness does not logically supervene on the physical (1996, pp. 94-95). Can we imagine a being (i.e., a zombie) that resembles a person in every physical aspect but lacks phenomenal experience? When this zombie steps on a thorn, its skin is pierced just like a human's would be, causing the nervous system (that works just like ours) to send messages to its zombie brain (substantially and functionally identical to a human brain), which causes the same motor responses a human might have, such as exclaiming or recoiling. The only thing missing in the event is the actual phenomenal experience of pain. If we can conceive of such an entity, then the physical processes occurring in the brain during consciousness cannot be said to rigidly designate the

[12] The inverted spectrum argument has been popularised in contemporary philosophy of mind due to the influence of Block and Sydney Shoemaker, but it is based on John Locke's thought experiment (Byrne, 2025).
[13] It is precisely this inverted spectrum argument, based on the works of Kripke and Block, that leads to the formulation of "the explanatory gap" (Levine, 1983, p. 356).
[14] "Psychological" here is taken from Chalmers' definition explained earlier in the chapter, referring to the functional aspects of consciousness (1996, pp. 11-31).

same thing as the phenomenal experience of consciousness (1996, p. 96). If physicalism is true, then zombies must not be possible.

Chalmers (1996, p. 96) states that the conceivability of zombies seems "obvious" (to him), but even if it is not, it is the premise of the argument which leads to the hard problem of consciousness: how do we account for phenomenal experience if consciousness is not supervenient on the physical?

> How can we explain why there is something it's like to entertain a mental image, or to experience an emotion? It is widely agreed that experience arises from a physical basis, but we have no good explanation of why and how it so arises. Why should physical processing give rise to a rich inner life at all? It seems objectively unreasonable that it should, and yet it does. If any problem qualifies as *the* problem of consciousness, it is this one (2003, p. 226).

That is what makes the hard problem of consciousness hard. It goes "beyond problems about performance and functions" (1995, p. 203). Even when we have described every functional process involved with consciousness, there seems to be something left to explain.

One objection often raised against the zombie argument questions the premise that *conceivability* suffices for *possibility*. For instance, the example is often used that it is perfectly conceivable to have a world in which water is not $H_2O$. In fact, before the composition of water was discovered in the 18$^{th}$ century, it was thought that water is a "simple, primitive substance" (Levine, 1998, p. 452). Now that we know that water has a molecular structure that is $H_2O$, it is no longer valid to use that conceivability as an indication of metaphysical possibility because chemical theory states that water is $H_2O$. In terms of consciousness, critics of the zombie argument claim that a question of conceivability is an epistemic one, whereas the question of consciousness is metaphysical (Kirk, 2023). According to this view, the conceivability of philosophical zombies tells us nothing about the metaphysical truth about phenomenal consciousness (Chalmers, 1996, p. 131). To explain why this argument is flawed, Chalmers distinguishes between "primary" and "secondary" senses of conceivability (1996, p. 132). Every concept has a primary intention (known a priori, like what we

mean when we think of water), and a secondary intention (known a posteriori, like the discovery that water is $H_2O$). In the case of consciousness, however, the subjective feel that has been described up to this point represents the primary meaning in its own right. P-consciousness is defined as something that has a “phenomenal feel” (1996, p. 133). In the “zombie world”, Zombie P-consciousness is still defined as something that has a “phenomenal feel” (1996, p. 133). The primary and secondary intentions of P-consciousness are therefore identical: "if something feels like a conscious experience, even in some counterfactual world, it is a conscious experience” (1996, p. 133).

Other responses to the zombie argument challenge Chalmers’ conclusion (1996, p. 96) that zombies are indeed conceivable. Dennett (1991, pp. 369-373) dismisses the very idea of subjective qualitative experience as an illusory phenomenon that results from convoluted literature and thought experiments. Chalmers (1995, p. 209; 1997, pp. 2-4) argues that ignoring or underestimating the hard problem goes no further towards explaining the mystery of subjective consciousness. To argue that this explanatory gap is non-existent or inconsequential simply “solves the problem by ducking the question” (1995, p. 206). Deflating the hard problem by using analogies to “easy problems” misses the point that Chalmers makes about P-consciousness's irreducibility to structure and function (1997, p. 4). At its most extreme, reductive physicalism can be taken to dismiss consciousness outright, which intuitively seems like an extreme position to take, considering that “[e]xperience is the most central and manifest aspect of our mental lives” (Chalmers, 1995, p. 206).

Frankish (2007, pp. 550-666) counters Chalmers’ conceivability argument with his own *anti-zombie argument for physicalism*. Broadly, if we can conceive of a purely physical zombie-world where our zombie duplicates do have P-consciousness, then “anti-zombies” are possible. Of course, the inverse conceivability argument is an imperfect parallel. Whereas phenomenal concepts are necessarily primary intentions in the zombie argument, as explained above, there is no a priori knowledge that phenomenal properties and microphysical properties co-refer. The anti-zombie argument is therefore vulnerable to the critique of the conceivability argument in a way that Chalmers’ argument is not.

The zombie thought experiment is particularly interesting to AI research and will be developed in Section 2.4. There are clear consequences for the “humanly-thinking” approach explained in the background section of Chapter 1. For if zombies are conceivable, even a perfect artificial reproduction of the physical/functional human neurological system will not necessarily bring about consciousness. However, if zombies are *not conceivable*, the hard problem of consciousness compromises physicalism as a complete metaphysical theory. The burden of proof shifts to the physicalist to show that functional outputs or structural frameworks cannot be recreated without subjective phenomenal experience.

## 2.3 Responses to the hard problem of consciousness

Thus far, the argument I have made is well established. The subjective nature of consciousness is a topic at the heart of contemporary philosophy of mind. In what follows, I consider some of the responses to the hard problem of consciousness.

There are some who are sceptical that there is even a problem to begin with. Daniel Dennett (2003, p. 10) likens questioning the nature of consciousness to finding out how magic is performed but being disappointed with the reductive explanation of the trick. Patricia and Paul Churchland (1981, p. 144) hold a similar view, arguing that the anti-functionalist argument relies on dualist misconceptions about consciousness. They represent a popular stance, arguing that the mysteries about the phenomenal nature of consciousness will be dissolved by further empirical investigation. According to strong reductionists and eliminativists like Papineau, Dennett, and the Churchlands, the hard problem is simply a combination of easy problems; if we have a physical explanation for all the easy problems, we will indeed have a complete account of consciousness.

Considering Jackson’s knowledge argument, Dennett (1991, p. 399) argues that if Mary truly did have access to all the relevant facts about visual experience, then she would gain no new information when she first sees colour. He says that the problem with Jackson’s thought experiment is in the way it is asked, as we are required to imagine Mary having access to all information about the physical world, which

surpasses the limits of our imagination (Dennett, 1991, p. 399). Dennett argues that if she indeed did know everything about the physical world, that would include the knowledge of redness even before she first sees colour (1991, pp. 399-400).

Paul Churchland (1996, pp. 219-220) uses the counterexample of “Blind Mary” in a thought experiment, who has all the information about the physical world but cannot see colours in a certain spectrum. According to his argument, Mary does not gain any more knowledge than Blind Mary when perceiving colour (1996, p. 220). Instead, Blind Mary simply lacks a certain “form” of the same knowledge that Mary has (1996, p. 219). This is where Chalmers’ Kripkean necessity is especially pertinent. One can know facts about Hesperus without knowing facts about Phosphorus, but when it is empirically discovered that both designate Venus, the facts about Hesperus must also be true about Phosphorus (Chalmers, 1996, p. 141). Churchland’s counterargument contradicts the notion of rigid designation between phenomenal experience and the physical world and therefore seems fallacious. For the sake of clarity, this simply means that, according to Chalmers’ interpretation of Kripke, Churchland’s argument that Blind Mary has equal knowledge to Mary is flawed. Intuitively, this seems to be a reasonable response, as both Dennett's and Churchland’s claims are essentially that Blind Mary would know what the phenomenal experience of red is, without ever having seen red (Jackson, 1986, pp. 294-295).

These eliminativist and illusionist approaches heavily leverage the epistemic vulnerability of our subjective experience. As critics rightly point out, human introspection and memory are highly fallible mechanisms (Dennett, 1979, p. 95). Dennett and the Churchlands argue that if our only access to phenomenal consciousness is through a faulty introspective mechanism, the phenomenon itself might merely be an illusion generated by the brain. This raises a critical epistemic question: what happens when the only way we can gain access to a particular phenomenon is through a faulty mechanism? However, it is important to keep the distinction between A-consciousness and P-consciousness in mind. Introspection belongs to the group of cognitive processes that Block (1995, pp. 230-235) identifies as the former. When introspection fails, it is a failure of the reporting mechanism. It does not prove the absence of the P-conscious feature. While introspection may be fallible regarding the specific content or external cause of a mental state, it cannot be

wrong about the bare fact that a subjective experience is happening. As Chalmers (1996, p. 133) notes regarding the primary intention of phenomenal concepts, a P-conscious state is defined by “a phenomenal feel”. The illusion of a P-conscious state therefore inherently constitutes a P-conscious state. The fallibility of introspection highlights the exact explanatory gap between our functional cognitive access and our underlying phenomenal reality that the hard problem of consciousness targets.

Physicalists who commit to explaining phenomenal consciousness within their framework are forced to embrace a restrained account of physicalism. Papineau (1993[2013], p. 32; 2024, p. 130) holds such a “naturalistic” view that sees mental events as both supervenient on the physical and irreducible. According to this theory, X has the function of Y when X exists because of carrying out function Y (1993[2013], p. 33). To achieve this, Papineau (1993[2013], p. 34) introduces “learning” as an alternative argument for supervenience through “selection” rather than reduction. By this view, pain (X) has the function of avoiding harm (Y). In other words, pain does not necessarily exist as a specific biochemical state, but rather it is attributed as the representationalist state of the function of avoiding pain. However, this still fails to bridge the explanatory gap as formulated by Jackson. Simply holding phenomenal-physical identity as brute does not explain consciousness. Chalmers (1996, pp. 136-138) argues that such a theory misuses the Kripkean *a posteriori* necessity explained in the previous section. Instead of responding to the hard problem of consciousness, “strong metaphysical necessity” shrouds objective experience in even more mystery by forcing it as a brute fact (1996, p. 138). By accepting that “consciousness” exists distinctly from physical function, consciousness is removed from the domain of physicalism. A physicalist theory leaves no room for the added element of experience.

J.J.C. Smart (1959, pp. 141-156) takes a similar approach to attempting to answer the question of consciousness without abandoning physicalism. To defend physicalism, he translates descriptions of experiences into “topic-neutral” language that avoids committing to either dualism or physicalism (Smart, 2017). Using the example of seeing an orange, Smart (1959, p. 149) suggests that instead of claiming "I see a yellowish-orange after-image" a more accurate topic-neutral report would be, "There is something going on which is going on when I have my eyes open, am awake, and there is an orange illuminated in good light in front of me." Because the event of

perceiving orange is not itself an orange object, it can simply be identical to a physical brain process. Again, Chalmers (1996, p. 360) points out that the act of perception that is being described concerns only A-consciousness and ignores the phenomenal aspects.

Some physicalists might agree that subjective consciousness exists, thereby accepting the hard problem, yet maintain that it will ultimately be solved through further reduction. Chalmers (1996, p. 167) calls them “type B” physicalists and explains:

> I have little difficulty in rejecting type B [i.e., physicalism that accepts the hard problem but claims it is solvable through reductive functionalism]. While it has the virtue of taking consciousness seriously, it relies on a metaphysics that is either incoherent or obscure, and one that is largely unmotivated; the main motivation is simply to avoid dualism at all costs (1996, p. 167).

If we accept that consciousness in the phenomenal sense is real, not just an illusory byproduct of neurobiological processes, and that the physical domain remains causally closed, then we are left with some form of interactionalist dualism (Chalmers, 1997, p. 44). Either experience exists outside the causal network epiphenomenally (i.e., dualism), or, as Chalmers (1997, pp. 302-308) suggests, we find a middle path between physicalism and dualism that allows us to continue meaningfully exploring consciousness. His proposal for *Naturalistic Dualism* will be discussed alongside panpsychism in Chapter 4.

First, considering what has been said about consciousness thus far, the connection to AI must still be established. This chapter has served as an introduction to the hard problem of consciousness. However, I will argue that the underlying theoretical metaphysics have direct implications for the real-world conceivability of AI technology that meets the threshold for proposed AGI. To achieve the overarching aims and objectives as outlined in the previous chapter, the following section explores artificial consciousness for AI.

## 2.4 Machine consciousness

There is a difference between the questions: “Can a machine think?” and “Could a machine be conscious?” (Chalmers, 1996, p. 313; Seth, 2025, pp. 1-2). The background section of this dissertation defines AGI as machines with human-level capacities in a general sense, as opposed to narrow AI. Searle (1980, p. 417) draws a similar, although not identical, demarcation between “strong” and “weak” AI. Weak AI views the computer as a “very powerful tool” that can process information and appear to have intelligence, whereas strong AI seeks to instantiate a mind (i.e., strong AI is a machine that is conscious) (1980, p. 417). Whereas narrow/weak AI is defined by its capability, Strong AI is defined by its being conscious. To be sure, the research question is specifically concerned with AGI in the strong sense. This section explores whether consciousness is theoretically artificially replicable, whereas Chapter 3 will consider whether AGI truly needs to be conscious.

The conceivability of conscious AI can be considered in similar terms that were used in the plausibility argument Chalmers makes for philosophical zombies. Whereas the thought experiments regarding inverted or absent conscious states were used to show that it is *conceivable* that consciousness does not supervene on the physical, Chalmers’ (1996, pp. 253-263) argument against the *possibility* of absent conscious states serves as a positive argument for the conceivability of artificial consciousness.[15] If we were to replace the biological neurons of a conscious human being with functionally isomorphous silicon chips, it would presumably not affect the conscious experience of the subject. However, if we continue to do this until the entire brain is replaced by silicon, would the subject lose consciousness at some arbitrary point? Or would consciousness gradually fade away as more neurons are replaced? The final possibility is that the subject’s consciousness will remain intact despite having an artificial brain (1996, p. 255). According to Chalmers (1996, p. 259), the “cognitive content” of the subject would be unchanged; therefore, artificial consciousness is conceivable.

Conversely, the idea of an unconscious system with such capacities (i.e., unconscious AGI) also poses problems for the physicalist account. Could we imagine an artificial

[15] Chalmers (1996, p. 253) calls it the “Fading Qualia” thought experiment to illustrate his argument against the natural possibility of “absent qualia”. I refrain from using the term qualia for the same reason as in the previous sections.

system with all the appropriate input, output, and internal states but lacking human consciousness? If these AI zombies — that would meet the description for AGI except that they are not conscious — are conceivable, physicalists need to account for why experience would not arise in such a system.[16] Dennett would have to accept that a system with such an architecture is equally as conscious as its human counterpart; this admission does not particularly trouble him, however, because consciousness — whether human or robot — is no more than a "theorist's fiction" in the first place, in his view (1991, p. 431). The Churchlands (1990, p. 32) would concur that, at a certain level, the machine will reach a state where there is no meaningful distinction between itself and a human counterpart. By their definition, if a system performs the correct processes, it can be considered conscious, whether it has a phenomenal experience of being or not. This means that such an advanced AI-zombie would, at some point, become conscious, so entailing that they would no longer be conceivable. If AI zombies are not conceivable, we need to be able to program consciousness to achieve AGI. In contrast, if such AI zombies are seen as conceivable – given our consistent functional advances in AI and generally perceived lack of progress towards programming consciousness, one might argue that this seems the likeliest – the question of why equal processes bring about consciousness in humans further highlights the tension in the functionalist's claim. Either way, the concept of AGI is troublesome for the physicalist argument.

If AI zombies are proven to be metaphysically possible in application and we design such an advanced system — i.e., if we achieve AGI without consciousness — there is even more of a burden on physicalists to account for the place of consciousness within a causally closed framework for the mind. Creating such systems would support Chalmers' argument that philosophical zombies are indeed conceivable, strengthening his case that the mental does not necessarily supervene on the physical. Recreating the physical system to which our own mental life is attributed, completely sans consciousness, would make it even more of a mystery to explain where phenomenal experience fits into our own inner physical workings. Still, what I have described as AI zombies would be a perfectly acceptable instantiation of AGI to the

[16] Or "Blockheads" after Block's (1981, pp. 6-7) hypothetical machine that can mimic human behaviour without consciousness.

physicalist. However, I argue that such a system can only ever be considered to be AI in the weak sense.

Searle (1980, pp. 417-420) demonstrates the problem of functional definitions of consciousness through a thought experiment.[17] We are asked to imagine a subject with a native grasp of the English language, but no knowledge of an example foreign language. The subject receives a set of English instructions (a program) for manipulating symbols in the foreign language (1980, p. 418). The English speaker is in a closed room and given a set of questions by speakers of the foreign language from outside the room (1980, p. 418). Because the interrogators cannot see the workings of the English speaker, they mistakenly assume that whoever is in the room understands their language because of the answers they receive (1980, p. 418). In fact, the English speaker is simply matching symbols based on a set of rules with absolutely no understanding of meaning (1980, p. 418). Searle (1980, p. 423) claims that machine systems operate in a similar manner to the subject in the thought experiment who naively produces accurate responses. Regardless of the success such a system might achieve in completing its given objectives, it cannot be thought to understand or think because it is mindlessly manipulating symbols with no comprehension or intentionality. Such systems are effectively purely computational "lookup tables" generating responses without any form of cognition (Millière & Buckner, 2024*a*, p. 2). In other words, a system cannot be judged to be conscious because of its functional success. Though the thought experiment in its original form specifically directs the argument against machine understanding or intentionality, "it is fairly clear that consciousness is at the root of the matter" (Chalmers, 1996, p. 322).

It is important to make a distinction here to avoid conflation. The *conceivability* of AI zombies is used to demonstrate that consciousness does not logically supervene on computational processes. Since AGI is not only a theoretical goal, the *possibility* of AI zombies is a completely separate question. This is a necessary note to keep in mind for the argument in Section 3.2.

[17] Since that is the example Searle (1980, pp. 417-420) used, it is well known as the "Chinese room" thought experiment. Recognising the needless and insensitive terminology, however, the dissertation will only be referring to it as "Searle's thought experiment".

Here, the confusion that stems from the Turing Test becomes pertinent. Searle's thought experiment shows that computational output cannot prove consciousness (Searle, 1980, p. 419). Turing (1950, pp. 433-435) recognises this and deliberately avoids attempting to do just that by explicitly changing the question from "Can machines think?" to: Can it behave in a way to make it indistinguishable from humans? He gives a pragmatic reply to what he calls *the argument from consciousness* (1950, pp. 444-445). Instead of attempting to argue that such functional processes could, in fact, bring about phenomenal consciousness (i.e., functionalism), Turing (1950, p. 445) makes the point that we have as little/much evidence of a machine being conscious as we do of other humans being conscious. Since, as Nagel (1974, pp. 435-449) argued, phenomenal experience is only known subjectively, Turing (1950, p. 444) concludes that "the most logical view to hold" is that the only way to really know that another person "thinks" is to be that person. This view leads naturally to solipsism.[18] According to Turing (1950, pp. 444-445), if one maintains that internal experience is the only proof of mind, then the existence of minds in others can never be truly verified. This is a fair assumption, given the problem of other minds.[19]

To navigate this dead-end and continue his research, Turing (1950, p. 444) suggests, "Instead of arguing continually over this point it is usual to have the polite convention that everyone thinks", proposing we extend the courtesy to any machine that shows the same behavioural evidence of consciousness (Russell & Norvig, 2022, p. 1036). This seems to be the point of confusion that has persisted in subsequent research into AI. AI research has remained interested in the behavioural capacities, functional architecture, and learning models without much consideration of philosophical questions about the nature of the mind. To be absolutely clear, this is no attack on Turing. I am simply making the point that "the view that it is abstract causal structure that is essential to mentality [i.e., functionalism] has been an implicit assumption of the AI research program since Turing" (Chalmers, 1992, p. 25). This is a necessary

[18] Solipsism is the philosophical position that the only thing that is certain to exist is one's own mind. This conclusion is usually reached by arguing that the enquiring mind must exist by virtue of questioning what exists (Avramides, 2023).

[19] That is, the epistemological issue of justifying the belief that other beings possess "thoughts, feelings and other mental attributes" (Avramides, 2023). The ongoing debate around attributing consciousness to animals or even the existence of other minds forewarns the challenges in ascribing consciousness to machine systems (Searle, 1980, p. 421; Chalmers, 2023).

foundation for an overview of approaches to AI that will be discussed in the following chapter.

If consciousness can be understood in terms of its function, artificial consciousness should be an easier goal to achieve than AI with equal functional capacities to humans (Chalmers, 2023; Millière & Buckner, 2024*b*, pp. 27-28). We accept, for instance, that there are animal species that have functionally simpler intellects than humans, but which display some level of consciousness (Chalmers, 2023). The bar for consciousness should therefore be lower than the bar for AGI. If computational functionalism holds and systems become conscious due to how they implement certain functions, rather than their actual physical makeup, it does not matter whether a system is biological or artificial (Chalmers, 1996, p. 320; Millière & Buckner, 2024*b*, p. 28). This is already a major concession for the physicalist, but it is a necessary one if the goal is conscious AI, because physicalism in its strongest sense "hasn't got a chance" to produce human-like consciousness in systems that are not human brains (Searle, 1980, p. 423).[20] More on this in the following chapter.

The previous sections have shown that such a reductive approach, at best, fails to provide a complete account of consciousness, and at worst, confuses us into adopting positions where we are forced into denying the existence or relevance of consciousness altogether. Physicalists need to deny the conceivability of zombies to avoid epiphenomenalism. If zombie worlds are conceivable, it seems to follow that P-consciousness plays no causal role. I have argued that zombies are conceivable, but that it points to a shortcoming in the physicalist account of the mind and not necessarily to epiphenomenalism. We can ask ourselves the question: as our artificial systems become more advanced, at what point do we concede that AI Zombies are real? And where do we fix the level where functionality equates to consciousness? [21] If functionalism holds, physicalists need to explain at what point a system suddenly becomes conscious. Once sufficient capability is reached across a wide enough range

[20] Searle might arguably be too harsh on token physicalism's chances of achieving conscious AI, saying "for unless you accept some form of dualism, the strong AI project hasn't got a chance". However, it is fair to say that conscious AI is a much harder challenge for the token physicalist. It requires not only the philosophical gap between consciousness and AI to be bridged, but also technological advances in engineering a substrate that would meet the requirements.

[21] This would lead to more problems in explaining why such a seemingly arbitrary cut-off point makes the difference between total darkness and a rich inner life within a system, whether human or artificial.

of “narrow” tasks, we will inevitably reach the point where a decision must be made regarding whether a machine that mimics human cognition well enough to pass the Turing test also possesses a phenomenal consciousness.

## 2.5 Conclusion

This chapter has spent time unpacking the hard problem of consciousness as the focal philosophical challenge at the heart of this dissertation. Chalmers’ formulation of the problem, illuminated by seminal thought experiments from Jackson, Kripke, Nagel, and Searle, challenges the assumption that the further examination of concurrent physical processes might lead to a complete reductive account of the mind. By engaging with strict physicalist objections to this challenge through Dennett, Frankish, the Churchlands, and Papineau’s varied approaches, the chapter has demonstrated that physicalism is unable to resolve the epistemic gap identified by the hard problem of consciousness.

I have argued that the phenomenal aspect of consciousness is a real and relevant part of consciousness, and that dismissing it goes no further towards explaining it. It has also been contended that physicalism may fundamentally be unable to describe that aspect of our mental experience. To establish the connection to AI, I briefly considered the conceivability of creating consciousness in artificial machines. This lays the foundation for the argument that follows in Chapter 3.

At this point in the dissertation, there is still only one (modest) commitment that needs to be made: consciousness exists (as more than a mere illusion). In the following chapter, Searle’s (1980, pp. 417-420) distinction between computation and understanding is used to argue that engineering consciousness – or at least recognising consciousness – in machine systems is needed for AGI. Since phenomenal consciousness has been distinguished from the reductive aspects in the overall phenomena of consciousness in the preceding sections, it must be argued that phenomenal consciousness is important for AGI. If this premise is accepted, it stands to reason that the hard problem of consciousness is also a problem for the conceivability of AGI.

# Chapter 3: Artificial General Intelligence

## 3.1 Introduction

This dissertation began by suggesting that the contemporary pursuit of AGI has reached a point where the underlying philosophical distinctions have become "continuous with the bleeding edge of engineering" (Koralus, 2024, p. 4). To explain how the metaphysical preamble of the previous chapters relates to Artificial Intelligence, this chapter examines the place of consciousness in realising AGI. In Chapter 2, I noted the close relationship between consciousness and intelligence and differentiated between weak and strong AI. While consciousness is not identical to intelligence, I argue that it plays a role if the target is *general* intelligence.

Of course, the definition of AGI within the relevant academic fields also remains a debated issue. In general, it is understood vaguely as AI capable of replicating the generality of human intelligence. Whether this involves the type of P-consciousness that concerns the hard problem of consciousness is another demarcating question that splits academia. Since Turing, researchers interested in developing AI have largely focused on the behavioural capacities of systems instead of their ontological substance. To be sure, this has been an effective strategy for creating narrow AI with ever-more impressive capabilities. Consciousness, especially in the phenomenal sense as has been defined in Chapter 2, is ignored. If one remains unconvinced of the hard problem of consciousness, this is to be expected. Why waste time arguing over machine consciousness if machine intelligence is the target? However, if one takes the hard problem of consciousness seriously and recognises the relevance of the phenomenal aspects of consciousness in the way we humans operate, then it becomes clear that any artificial system requires some form of consciousness to fully emulate the human mind.

After making the argument that AGI requires artificial consciousness (and specifically the phenomenal aspects thereof), the methods researchers are employing to achieve the goals of AI are considered through Russell and Norvig's AIMA quadrant, and I argue that the phenomenal aspects of consciousness have been neglected. I will

summarise the different philosophical approaches to explaining consciousness discussed thus far, along with the implications for conscious AI. While the rapid advancement and success of narrow AI have shifted the boundaries of what we envision for technology, the path toward a system that can process human-level general intelligence remains obscured by what John Haugeland (1985, p. 247) refers to as a "theoretical time bomb, ticking away in the belly of AI": the unresolved problem of consciousness.

## 3.2 Consciousness for AGI

Having shown that consciousness is conceptually distinct from intelligence, the argument needs to be made that it is necessary for AGI. Searle's thought experiment shows that there is a difference between computing and understanding, but I have yet to show that understanding is required for AGI. This section is dedicated to making the argument that phenomenal consciousness is indeed required for AGI.

As noted, the question of whether consciousness is necessary to achieve AGI is subject to debate. Those who think we are able to create AGI without consciousness point to the distinction between the functional capacities that define intelligence and the subjective experience of consciousness (Seth, 2025, p. 34). In fact, the consensus amongst AI researchers appears to be that it is the ability to process information, not the ability to feel or experience the information, that is most important. Section 3.4 considers the different approaches to AI and shows that phenomenal consciousness is not a priority in contemporary research.

The most straightforward way to make the case that AGI should be understood as artificial systems with capacities including consciousness that match humans on a general level is to point at exactly the same separation of consciousness and intelligence. If AGI is taken to be conceptually different from narrow AI, what separates a system that is capable in a wide array of specified tasks from one that is generally intelligent? Whereas narrow AI has faced relatively little difficulty attempting to replicate functional elements of intelligence, creating AGI appears to be a completely different challenge. Comparably, since consciousness and intelligence, though related,

are not identical, one seems easier to explain reductively than the other. I argue that these two problems are intertwined. The remarkable rate of progress in AI technology makes it increasingly difficult to imagine that machines would not be able to simulate many aspects of human behaviour at some point. If this suffices for AGI, there is not much of an argument to be made. However, if AGI is taken to be conceptually different from advanced forms of narrow AI, replicating certain (however many) elements of human behaviour does not seem sufficient for AGI. Searle's (1980, pp. 417-420) distinction between syntax and semantics is therefore used to argue that engineering consciousness in machine systems is needed for AGI.

Phenomenal consciousness has been distinguished from the reductive aspects of the overall phenomena of consciousness. Accordingly, I argue that it holds its own intrinsic value. While it is conceivable for the processes involved in consciousness to occur without accompanying phenomenal qualities in theory (which is the basis for the argument I have made against physicalism), it remains "a fact about the world" that they are co-occurring (Chalmers, 1996, p. 18). I argue that it seems clear that phenomenal consciousness plays a necessary role in human mentality. Aspects like "flexible control" of thought, "social coordination", "integrated representation" of a unified reality, "informational access", "freedom of will" and "intrinsic motivation" are suggested to all be some of the possible functions of consciousness (Seth, 2009, pp. 286-292; Van Gulick, 2025). It is not difficult to imagine that P-consciousness is involved to some degree.

If the role of consciousness in these cases is constitutive, it is self-evident that AGI needs to instantiate P-consciousness. If it is said to be contingent, then consciousness does not supervene on the physical, and zombies are conceivable, leaving us with the threat of epiphenomenalism again. Even so, since our only access to these mental functions is through our own phenomenal experience, it is unclear how we could program a system with similar capabilities without a comparable level of consciousness. What would it mean to create a program that can make decisions to avoid suffering without the capacity to experience or understand suffering? Can Mary, the scientist in the black-and-white room, make a fully informed decision about which colours to use when she redecorates her home, based only on the physical facts she has learned about colour perception? The frame problem, discussed in the following

section, further argues for the importance of the human ability to make judgements on context relevance that seems crucial for general competence. It seems like any meaningful definition of AGI would be strong AGI. Considering then that we take consciousness to have epistemological value as well as causal status without being supervenient on the physical, I argue that any system that meets the requirements of AGI would need functioning P-consciousness. It is important to emphasise that my argument does not affect the conceivable success of narrow AI. All evidence seems to suggest that the current approaches (discussed in the following section) can bring about highly effective/dangerous systems based on a functionalist understanding of the mind by recreating specific elements of intelligence. Understanding that phenomenal consciousness plays a role in the adaptability and inference capabilities that differentiate AGI from narrow AI simply means that current approaches are limited to the latter.

This might appear to be an objectionable tension in the argument I have made. I have used the conceivability of AI zombies to challenge functionalism. Now I am suggesting that AGI requires consciousness, thereby denying AI zombies. My reasoning for associating consciousness with AGI is different from that of the functionalist. Functionalists claim that AGI requires successfully simulating the underlying mechanisms of phenomenal consciousness; however, these processes (e.g., awareness, learning and attention) can be identified with the A-conscious elements of the mind (Cea, 2025, p. 68). In contrast, I argue that comprehension and meaning rely entirely on the subjective state that accompanies them. For a system to act autonomously and creatively in a novel environment, it requires phenomenal consciousness. Instead of taking the problem of consciousness to be a problem of cognitive function, I propose “it should be seen as a problem of phenomenal experience” (Haikonen, 2013, p. 20).

Axel Cleeremans and Catherine Tallon-Baudry (2022, pp. 1-11) make a corroborating argument for the relevance of phenomenal consciousness. They use examples of the intrinsic value of subjective experience to demonstrate the point. Emotions and decision-making are often, and especially in the context of machine learning, argued to be relatively simple to recreate without reference to the subjective elements involved. Usually, this involves simply assigning a reward to a preferred outcome in the case of

decision-making or by referring to additional dimensions to characterise emotion (2022, pp. 3-4). Cleeremans and Tallon-Baudry (2022, p. 3) argue that external (i.e., programmed) value cannot have intrinsic value because "the key for a reward to have 'intrinsic' value is precisely subjective experience, for in its absence, there is nobody for whom the reward means anything." They also make a case that sensory experience is "always associated with an intrinsic subjective value" (2022, p. 4). This leads to the conclusion that P-consciousness plays a valuable role in flexible and creative behaviour, and internal drive and motivation (2022, pp. 4-6). They concur that the hard problem of consciousness means that phenomenal consciousness lies outside the scope of explanations that consider the performance of functions, but that subjective experience nonetheless has its own inherent functional value (2022, p. 8). Any conception of AGI would surely include faculties like creativity, curiosity, doubt, inspiration, empathy, fear, or drive to be sufficiently broad as to qualify as AGI. If phenomenal consciousness plays a functional role in those faculties, it follows that AGI needs to be conscious.

Consciousness inessentialists, those who see no logical or metaphysical reason for phenomenal consciousness, contend that philosophical zombies are not only conceivable but that it is "difficult to explain why we are not" zombies ourselves (Flanagan & Polger, 1995, p. 321). This is clearly incompatible with most physicalist approaches that are forced to rule out the conceivability of zombies by claiming a metaphysically necessary link between consciousness and the underlying physical processes (Seth, 2009, p. 282). Physicalists would have to agree, therefore, that consciousness plays a necessary role in the different behavioural capacities that collectively define *general intelligence* and distinguish it from specific elements of intelligence that are the domain of narrow AI.

However, even if it is not accepted that consciousness is required for AGI, it still matters whether AI is conscious or not (Chalmers, 2023, p. 3).

> Conscious systems have moral status. If fish are conscious, it matters how we treat them. They're within the moral circle. If at some point AI systems become conscious, they'll also be within the moral circle, and it will matter how we treat them. More generally, conscious AI will be a step on the path to human level

> artificial general intelligence. It [could] be a major step that we shouldn't take unreflectively or unknowingly (2023, p. 4).

If one disagrees with the premise that machine consciousness is needed to achieve AGI, there are enough moral considerations and safety concerns related to the human ability to recognise consciousness in artificial systems — whether engineered or emergent — to warrant an interest in the research question.

The following section examines the prevailing methodologies of AI research as categorised by Russell and Norvig's (2022, pp. 19-22) AIMA quadrant. The aim here is to highlight how current approaches ignore and bypass the hard problem of consciousness. I evaluate a taxonomy of existing theories of consciousness to demonstrate the effect of the underlying metaphysical assumptions on the conceivability of conscious AI.

## 3.3 Current approaches to AI

Russell and Norvig (2022, pp. 19-22) differentiate between a quadrant of possible combinations of methodological approaches to developing AI. The four possible goals for AI — known as the AIMA quadrant — are: (1) "Acting Humanly: the Turing Test approach"; (2) "Thinking Humanly: the cognitive modelling approach"; (3) "Thinking Rationally: the 'Laws of thought' approach"; and finally, (4) "Acting Rationally: the rational agent approach" (2022, pp. 19-22). These quadrants provide a practical framework of possible approaches to AI that are not mutually exclusive but rather serve as complementary lenses for research. However, I contend that each quadrant relies on functionalist intuitions about the mind that identify intelligence with process or output. If AI zombies are conceivable, such an approach may well one day achieve AGI. However, I have argued that the hard problem of consciousness shows that any theory of the mind which excludes the integral phenomenal aspects could not provide a blueprint for instantiating consciousness in artificial systems. Having also made the argument that a system would need to be P-conscious to qualify as AGI, it follows that the hard problem of consciousness demonstrates why current AI systems cannot be conscious and therefore AGI (as I have defined it) cannot be achieved through the prevalent approaches.

To substantiate this argument, we need to consider the respective approaches seriatim. For an artificial system to recreate human behaviour, or “act humanly”, it would need a complex mixture of capabilities (2022, p. 20). Such a machine would need to be able to communicate with humans, adapt to circumstances, and detect patterns (2022, p. 20). It would also need to be able to store new information and reason to draw conclusions from said information (2022, p. 20). Turing (1950, pp. 433-435) proposed making use of *The Imitation Game* to test for intelligence in a system by gauging its ability to “think” by observing its behaviour. The goal of the system in the imitation game is to compete against a human opponent, answering a series of questions and attempting to convince its human interrogator that it too is a human participant (1950, pp. 433-434). Crucially, the imitation game is not interested in whether such a machine thinks like a human or if it *thinks* at all, which is why it exemplifies this approach to AI seeking to create systems that can convincingly *act humanly*. Recalling the discussion in the opening two chapters, it becomes clear that this approach does not consider the phenomenal aspects of consciousness by its very design.

The second approach to creating systems that “think humanly” focuses on modelling an artificial mind based on a precise theory of the mind that would need to be developed through extensive brain imaging and rigorous psychological experiments (Russell & Norvig, 2022, p. 20). The aim is to correlate brain imaging data with psychological fundamentals to create a model for machine-learning systems. The irreducibility of phenomenal consciousness challenges the underlying assumptions of this approach. Chalmers (1996, p. 94) uses the conceivability of “philosophical zombies” to argue that mapping out the neural processes occurring in the brain while experiencing consciousness goes no further towards explaining the phenomenal character of the experience. We have neither empirical evidence nor a logically convincing argument to suggest that synthetically recreating these physical processes would magically conjure up consciousness.

In contrast, the “logicist tradition within artificial intelligence,” seeking to create systems that “think rationally,” is rooted in Aristotelian logic rather than the physical processes that occur in the mind (Russell & Norvig, 2022, p. 21). According to this approach, the

goal is to find or create a strict set of “logical” rules to emulate human thought (2022, p. 21). Systems that “think rationally” are computationally programmed to have formal logical parameters. Such an approach makes intuitive sense for narrow tasks; however, real-world application rarely follows formal logic, and the uncertainties and unknowns of reality make it exceedingly difficult to create a complete set of rules that could account for all the variables (2022, p. 21). This approach most closely resembles what Haugeland (1985, p. 112) termed “Good Old Fashioned Artificial Intelligence” (GOFAI). While this works for a system with a narrow task with closed parameters, like playing chess, it does not engage with the phenomenal aspect of consciousness at all. In terms of AGI, the logic-based approach is challenged by the *frame problem* (Fodor, 1983, p. 114; Shanahan, 2016). The frame problem asks how a system relying on explicitly stored, sentence-like representations of the world can successfully update its beliefs in an open-ended environment. If an artificial system must reconsider its entire database of information to determine the consequences of a single action, it faces computational intractability. To avoid this, the system must rely on "relevance." However, defining what is relevant is entirely context-dependent, leading the AI into what Hubert Dreyfus (1992, p. 133) termed an "infinite regress of planning of contexts." The epistemological frame problem thus exposes a severe limitation of computational functionalism: it cannot tractably capture holistic, context-sensitive relevance using formal logic alone. I have made the argument that AGI would need to be able to understand context and that such understanding presupposes phenomenal capacity.

Finally, the “acting rationally” approach can be considered the most prevalent in AI research, simply because it is the broadest in scope (2022, p. 22). For AI to be a rational agent, Russell and Norvig (2022, pp. 21-22) stipulate that “agents” are expected to “operate autonomously, perceive their environment, persist over a prolonged time period, adapt to change, and create and pursue goals”. Such a system is rational when it “acts so as to achieve the best outcome or, when there is uncertainty, the best expected outcome” (2022, p. 22). This offers two advantages over the alternatives discussed. First, contrary to the approaches that focus on the pure logic involved in reason, logical inference is seen as merely one of several ways to achieve rationality. Secondly, it lends itself more easily to scientific progress. Since “rationality” in this sense is well-defined, researchers can often reverse-engineer systems that demonstrably meet this standard (2022, p. 22). The rational agent approach also

generally subsumes the acting humanly approach as “all the skills needed for the Turing test also allow an agent to act rationally” (2022, p. 22).

To understand these separate quadrants, consider the familiar examples of LLMs like ChatGPT, DeepSeek, or Gemini. These systems are pretrained on massive, multimodal datasets, using self-supervised learning to make predictions (Gemini Team, 2024, pp. 1-4; Jin, Tangsrivimol, Darzi, Hassan Virk, Wang, Egger, Hacking, Glicksberg, Strauss & Krittanawong, 2025, pp. 1-12). Following this initial training, the models are further trained by incorporating human evaluations of their test results to inform future predictions in what is known as *reinforcement learning from human feedback* (RLHF) (Jin et al., 2025, p. 2). The process is repeated to align their outputs with their programmed objectives. When operated, the system responds to a user prompt (input) by predicting a sequence of words that form a coherent response (output) (2025, pp. 4-6). While each model makes use of distinct datasets and fine-tuning methods, their underlying operational logic remains synonymous with the statistical pattern recognition characteristic of modern Large Language Models (LLMs).

On the AIMA quadrant, such models fall most naturally into the “acting rationally” and “acting humanly” sections. They are specifically designed to mimic human conversation and “sound” human (Jin et al., 2025, pp. 2-3). They are also compatible with the “acting rationally” approach, as they are trained to optimise goal satisfaction within the context of their programming and prompts. However, they lack key elements of being a full rational agent, such as autonomous goals, planning, or beliefs (Russell & Norvig, 2022, p. 22). So, while a system might generate a rational response in the sense that it maximises a conversational reward, it lacks the internal states necessary for genuine agency. They are less compatible with the “thinking rationally” approach because they are LLMs, making use of statistical pattern recognition, not formal logic, to produce an output. Nevertheless, some of the most recent updates have allowed models to mimic logical chains (OpenAI, 2025; Lightman, Kosaraju, Burda, Edwards, Baker, Lee, Leike, Schulman, Sutskever & Cobbe, 2024, pp. 1-29). It might, therefore, be argued that it is “thinking rationally” (when activating the “reason” option on ChatGPT, it will produce a chain-of-thought response that approximates logical inference). It is important to understand, though, that such a model still operates on probabilistic prediction based on datasets and lacks genuine symbolic or logical

reasoning (Jin et al., 2025, pp. 3-6). For this reason, such models do not belong to the "thinking humanly" quadrant on the AIMA index because human brains do not run through massive datasets using programmed training to predict goal-oriented responses; human cognition involves logical inference and understanding that do not mirror the programmed statistical training of an LLM.

Considering the arguments put forth by Searle, Jackson and Chalmers in the previous chapters, however, the AIMA quadrant remains silent on the phenomenal aspect of consciousness (Bringsjord & Govindarajulu, 2025). This is by design. I have explained that Turing chose to avoid the question of consciousness to progress with the research he was working on (Russell & Norvig, 2022, p. 1035).

> I do not wish to give the impression that I think there is no mystery about consciousness. There is, for instance, something of a paradox connected with any attempt to localise it. But I do not think these mysteries necessarily need to be solved before we can answer the question with which we are concerned in this paper (Turing, 1950, p. 447).

Haugeland (1985, p. 6) describes Turing as "annoyed by fruitless disputes over word meanings". To get past the "armchair philosophizing about what we mean by 'think' and 'intelligent'", Turing proposed the Imitation Game (Turing, 1950, pp. 433-435; Haugeland, 1985, p. 6; Russell & Norvig, 2022, p. 1035). If the goal is narrow AI, then this approach is reasonable and has been extremely successful thus far. Yet, if the goal is AGI, I have used Jackson's (1982, p. 130) knowledge argument and Searle's (1980, pp. 417-420) thought experiment to demonstrate that such a physicalist/functional account of consciousness fails to provide a comprehensive description of conscious experience and computation is not sufficient for understanding. There can be no Turing-like test to definitively measure the existence of phenomenal experience in artificial systems as certain as there can be no sensor to measure and confirm the consciousness of other human and non-human animal minds (Russell & Norvig, 2022, p. 1037). The hard problem of consciousness distinguishes phenomenal from functional consciousness and exposes the ontological gap in a physicalist account of the mind. I argue that this gap represents a theoretical blind spot in AI research and that we cannot create AGI without addressing the

shortcomings in our ontological framework. The hard problem of consciousness is, therefore, also a problem for AGI.

I have argued that any approach that excludes the phenomenal nature of consciousness will inevitably hit the “barrier of meaning” identified by Mitchell (2019, p. 51). There is a significant ontological gap in the physicalist assumptions underpinning the AIMA quadrants, which exclude phenomenal experience from the start. In fact, the success of systems like ChatGPT, DeepSeek, and Gemini in simulating human output without the emergence of consciousness places the burden of proof on physicalists to explain why subjective experience does not accompany the functional capacities of such systems, when that is what seems to be the case in humans. The following section overviews the existing theories of consciousness that have been discussed and considers whether artificial consciousness is conceivable given the respective explanations of human consciousness.

## 3.4 Theories of consciousness for artificial consciousness

As mentioned in Chapter 1, the “landscape” of competing theories of consciousness has been developing since long before the question of artificial consciousness. Robert Kuhn (2024, 28-35; 2026) differentiates between no fewer than 300 theories of consciousness. Instead of attempting to consider each of the varying and overlapping theories individually, this section will provide a summary of the main theories of consciousness in philosophy discussed thus far and consider the implications for the prospect of AI consciousness.

Cartesian dualism was used to illustrate the mind-body problem in the background section. To summarise, substance dualists believe that there are two distinct substances (mental and physical) that make up reality (Robinson, 2023). In terms of AI, dualism neither denies nor confirms the possibility of AI consciousness. However, if consciousness is distinct from the material realm, it lies outside the reach of what we can create artificially. The dualist may argue that AI consciousness might correlate with the physical substance of the AI system in the same way that human consciousness correlates with the brain. Considering the challenges levelled at

dualism to explain precisely that interaction between the mental and material, however, it seems to lack the explanatory value we require for a theory of consciousness for AI. Chalmers' (1996, pp. 124-129) property dualism is discussed in more detail alongside panpsychism in the next chapter as one of the alternative approaches that are not undermined by the hard problem of consciousness.

In contrast, physicalism is monistic about the substance of reality (Stoljar, 2024). As a broad category referring to theories that take consciousness to depend on physical states, it can be subdivided into several varying interpretations and definitions. Eliminativism and illusionism have been discussed in depth as the strictest forms of physicalism. To summarise, eliminativists claim that consciousness can be explained away by describing all the physical processes involved, while illusionists dismiss the phenomenal aspects of consciousness as fictitious side effects of the aforementioned physical processes. As explained using the concept of Zombie AI, strong reductionists see neither the possibility of nor the need for artificial consciousness. I have argued that such a reductive approach leads to unreasonable conclusions and is based on the fact that we only have subjective access to the particular phenomenal aspects in question (Chalmers, 1996, p. 102).

Computational functionalism defines consciousness by the computational role it plays within the larger system. As discussed, computational functionalism seems to be the most amenable to the conceivability of artificial consciousness. In fact, Putnam (1967, pp. 158-167) advances the view specifically to oppose the type-identity theories that would make machine consciousness unrealisable. In the previous chapter, however, Searle's thought experiment was used to argue that computational functionalism alone cannot give rise to a mind. Still, if conscious AI is the goal, any physicalist research program would need to defend computational functionalism, as it seems the only viable approach (Seth, 2025, p. 9).

Another common physicalist view of the mind that has not been discussed here yet is known as *biological naturalism*. This can be understood as "the idea that consciousness is a property of only, but not necessarily all, living systems" (Seth, 2025, p. 7). In essence, proponents of biological naturalism, like Searle (2002, p. 60), take consciousness to be "a biological process like digestion, photosynthesis, or the

secretion of bile." By definition, such a view is sceptical about the conceivability of consciousness in artificial systems, although Seth (2025, p. 25) argues that "real artificial consciousness" may be achieved by creating "machines that are also in some relevant sense alive." Chalmers (1996, pp. 375-376) notes that such an approach runs into the same problems as other forms of physicalism when challenged to explain consciousness in that it seems equally conceivable to have a zombie without P-consciousness that is *biologically* identical to a human counterpart.

Similarly, other physicalists take phenomenal consciousness to be an emergent property in a weaker way than the supervenience or functional claims. Chalmers (2006, p. 247) defines weak emergence as follows:

> We can say that a high-level phenomenon is weakly emergent with respect to a low-level domain when the high-level phenomenon arises from the low-level domain, but truths concerning that phenomenon are unexpected given the principles governing the low-level domain.

For example, intelligence would be an unexpected emergent property strictly in terms of genes (2006, pp. 253-254). However, taken from "the level of the organism" (i.e., considering a human, rather than just a brain), intelligence becomes less emergent because it is "no longer unexpected in the same way" (2006, pp. 253-254). In terms of consciousness, this observer-dependent approach does little to answer the metaphysical questions about the mind (Bedau, 1997, p. 399; Chalmers, 2006, p. 250). Such an approach to consciousness does theoretically allow for artificial consciousness, provided that the artificial system can perfectly recreate the physical organisational complexity that underlies conscious experience in some living beings. However, I have used Chalmers' arguments to illustrate that an epistemological explanation like weak emergence can not describe the nature of consciousness. This leads Chalmers (1996, pp. 124-129; 2006, pp. 249-252) to argue for a stronger sense of emergence. I consider Chalmers' proposal in Chapter 4, which is where the second part of the research question (namely, panpsychism as a possible alternative framework to ground AGI) comes to the fore. Strawson (2006, p. 53) uses an argument against strong emergence to make the case for panpsychism.

There are clear ethical implications accompanying either a positive or negative assertion on the conceivability of conscious machines. Humans standardly serve as the typical example of a moral agent in that “an entity is a moral agent if and only if it has obligations, and it is a moral patient if and only if moral agents have obligations towards it” (Müller, 2026). To qualify as a moral agent, an entity needs access to “knowledge about the world”, the “ability to act in the world”, the ability to “reason” and reflect, and crucially, “the possibility to be influenced by suffering or joy, technically phenomenal states with valence, or sentience” (Müller, 2026).

While the physicalist paradigm has advanced our descriptive and predictive understanding of intelligence, it has done so by abstracting away the very phenomenon that this study proposes to be a fundamental aspect of the human mind: consciousness. Consequently, to move from simulation to instantiation, I suggest that AGI research would need to transition away from purely functional benchmarks and toward a framework that treats consciousness as a fundamental feature of reality, such as the panpsychist alternatives explored in the following section. I will argue that panpsychism need not reject physicalism outright. Rather, it is a way of expanding the physicalist ontology to include phenomenal properties on a fundamental level. In terms of AGI, this means that systems would need to instantiate the same fundamental properties that underlie human consciousness. AGI is therefore still conceivable under such a framework, but there are much stricter requirements than proposed under popular functionalist and representationalist frameworks. This might seem counterintuitive. After all, if panpsychism claims consciousness is fundamental, then machines must already be fundamentally conscious. In the following chapter, I explore what such panpsychist AGI requires.

## 3.5 Conclusion

What does this all mean for the conceivability of AGI? By considering the different types of available theories of consciousness through this lens, it becomes clear that the AGI project poses not only a technical challenge but an ontological one. Following from the argument that began in the previous chapter, the intrinsic value of subjective experience appears to be an essential aspect of many of the capacities we associate

with general intelligence. The consequence for AI is that it seems that phenomenal consciousness is a necessary condition for AGI. The inability of physicalism to explain how and why phenomenal consciousness occurs suggests that this is an insurmountable challenge and that AGI cannot be achieved given a physicalist framework (even if it may be argued to be conceivable). The conceivability of AI zombies was used to illustrate the double-edged nature of the problem of AGI for physicalism: if consciousness is relevant to the capacities required for AGI, we need to be able to explain it through structure, yet the hard problem persists. If consciousness is irrelevant, pressure mounts on physicalists to account for its existence in the human mind.

A review of the current approaches to AI exposed a blind spot in methodology that disregards the phenomenal aspects of consciousness. I suggest that this is a symptom of the limitations of physicalist assumptions that inform the prevailing approach to engineering the mind, as revealed by the hard problem of consciousness. The crucial insight is that the most pressing challenge in creating AGI is the difficulty of programming phenomenal consciousness. By summarising the conceivability of conscious AI given the theories of consciousness that have been discussed up to this point, I show that our ontological commitments have real implications for the AI project.

Returning then to the line of reasoning that leads to the hard problem of consciousness, we must see Chalmers' argument through to completion. If the physical facts about brain states do not rigidly designate the same thing as the phenomenal experience of consciousness, it does not matter whether we are considering a brain or a computer, because functional or physical identity would not necessitate consciousness. It would also mean that we are left unable to attribute it to (much less engineer it in) ourselves or anyone/anything else within a physicalist framework because it would, by definition, fall outside of such a framework. Here, the second part of the research question regarding alternative frameworks for understanding the mind becomes pertinent. Chapter 4 focuses on that particular aspect. By now, the first part, namely, the consequences of the hard problem of consciousness for achieving AGI within a physicalist framework, ought to be clearer.

# Chapter 4: Towards a complete account of consciousness

## 4.1 Introduction

The hard problem of consciousness challenges the physicalist account of proposed strong AGI by arguing that merely recreating the functions or processes of the brain would not instantiate consciousness. Since I have argued that consciousness plays an important functional role in the human capacities that are set as the goal for AGI, any account that does not give a full description of the mind that includes consciousness would be an insufficient blueprint. Given the gap between a physical explanation and our lived experience, any adequate metaphysical framework must address this problem without collapsing into reductionism or epiphenomenalism. This chapter evaluates two options for such a framework, namely, naturalistic dualism and Russellian monism, with the aim of introducing panpsychism as a viable alternative to physicalism.

First, Chalmers' (1996, p. 124), *naturalistic dualism* is considered as one approach to bridging the gap. The tension with this approach lies in whether it can avoid epiphenomenalism (if epiphenomenalism does indeed need to be avoided). Chalmers' conclusion is that the ontological implications of phenomenal consciousness call for a broadening of our metaphysical framework (1996, p. 171). This requires a shift from a purely physicalist account to one that includes phenomenal consciousness in its scope.

The second section introduces the Russellian concept of *quiddities* – "properties that underlie structural and dynamic properties described by fundamental physics" (Alter & Pereboom, 2023, p. 1410). Describing the nature of these properties leads naturally to one of the most persuasive justifications of panpsychism, which is then introduced in the following section. Considering the cases for panpsychism, particularly in its Russellian form, the third section makes the argument that panpsychism offers the most complete and parsimonious account of the mind. Yet, a serious challenge exists

in various forms of the combination problem that still needs to be addressed if panpsychism is to be accepted as an alternative framework for proposed AGI. Thus, despite the interdisciplinary nature of the research question, this chapter necessarily concentrates on the philosophy of mind.

Having considered the first half of the research question in the opening chapters, namely, the hard problem of consciousness for AGI, this chapter moves on to the second part: Panpsychism as an alternative framework for understanding phenomenally conscious AGI.

## 4.2 Naturalistic dualism

The perceived challenges faced by physicalism lead to the need for an "additional component" (Chalmers, 1996, p. 126). One way to approach this is by placing phenomenal experience at a fundamental level in our account of the natural world.[22] This idea will be explored in the following section. The other approach, which Chalmers (1996, pp. 126-129) ultimately endorses, is to derive phenomenal experience from another class of novel fundamental properties.[23] In essence, this is a form of dualism as it argues that "there are both physical and nonphysical features of the world" (1996, p. 124).

There is an important distinction between this naturalistic dualism and the Cartesian substance dualism discussed in Chapter 1. Whereas substance dualism claims a metaphysical divide between mind and matter, Chalmers (1996, p. 125) describes his approach as a type of property dualism in its strongest sense. Further clarification is required, however, since property dualism is often taken to be a nonreductive form of physicalism (H. Robinson, 2023). For instance, Donald Davidson ([1970]2001, pp. 207-225) introduces Anomalous Monism, which is also defined as property dualism because mental events "resist capture in the nomological net of physical theory."

[22] That is, panpsychism. The Anti-Emergence Argument for panpsychism is closely tied to Chalmers' argument against physicalism explored in the previous chapter.
[23] Chalmers (1996, pp. 154, 299) admits that the alternative, panpsychist approach solves the supervenience problem that physicalism faces when accounting for consciousness. He prefers property dualism mainly to avoid having to commit to the universality of experience in a panpsychist framework (1996, pp. 298-299, 357).

Davidson ([1970]2001, p. 214) proposes "that mental characteristics are in some sense dependent, or supervenient, on physical characteristics" without being identical to them. This is perfectly compatible with non-reductive physicalism. Chalmers' (1996, p 125) claim is stronger: "Consciousness arises from a physical substrate in virtue of certain contingent laws of nature, which are not themselves implied by physical laws." Conscious experience is therefore ontologically distinct from the physical properties that accompany it.[24] Davidson's *predicate dualism* falls under the physicalist camp as far as the hard problem of consciousness is concerned.

While Chalmers is explicitly sceptical about physicalism, his view is clearly not a straightforward dualism. The defining element of Chalmers' (1996, p. 128) argument is that it is also "naturalistic", by which he means "it can explain consciousness in terms of basic natural laws". He uses the analogy of James Clerk Maxwell, who introduced "new fundamental electromagnetic laws" necessary to explain the newly discovered property of charge, which previously did not exist under Newtonian physics (1995, p. 209; 1996, p. 127). The discovery of new electromagnetic phenomena did not undermine physics; it simply called for its expansion. "To explain electromagnetism, the ontology of physics had to be expanded" (1995, p. 209). Similarly, naturalistic dualism does not undermine physics; instead, it is meant to explain the nature of the phenomenon that lies outside the scope of physics. The "psychophysical laws" bridge the physical and the phenomenal (1996, p. 170). Chalmers (1996, p. 171) maintains that "this dualism, then, requires us to give up little that is important about our current scientific worldview. It merely requires us to give up a dogma." With this naturalistic interpretation, Chalmers' (1996, p. 162) suggestion is that consciousness does supervene on the physical "naturally" (i.e., through natural laws), not metaphysically.

By placing consciousness outside the realm of physical properties, Chalmers (1996, p. 128) also avoids being undermined by the empirical success of science based on "physical explanations of behaviour and of other physical phenomena". However,

[24] Chalmers (2010, p. 139) explains that he intends to remain neutral on the question of substance, though his argument in *The Conscious Mind* (1996, p. 126) is explicitly not physicalist in that "the existence of further contingent facts [i.e., natural laws] over and above the physical facts is a significant enough modification to the received materialist world view to deserve a different label."

accepting the physical world as causally closed, Chalmers commits to at least a degree of epiphenomenalism (1996, pp. 150-159).

> If we assume that the physical world is causally closed and that consciousness causes some physical events, then it follows under certain natural assumptions about causation that consciousness must supervene logically (or metaphysically) on the physical. If so, then given that the physical world is causally closed, the mere natural supervenience of consciousness implies that consciousness is epiphenomenal (1996, p. 150).

Of course, critics of Chalmers' argument investigated in the previous chapter, like Dennett and the Churchlands, deny that such laws could exist independent of the physical systems they govern. These concerns have been dealt with in Chapter 2. The strictly reductive accounts have been shown to ultimately reject consciousness altogether or dismiss it as an illusory byproduct.

The most intuitive reason to want to avoid epiphenomenalism is that it seems completely self-evident that our conscious experience affects our physical actions (Chalmers, 1996, p. 159; W. Robinson, 2023). Of course, what appears to be self-evident is not always undeniable. Other arguments against epiphenomenalism are more damning. Most relevant to this dissertation are the arguments relating to the knowledge of the existence of our own mind, as well as that of other minds (W. Robinson, 2023).[25] If epiphenomenalism is true, then consciousness has no causal influence on anything, including our beliefs, experience, or knowledge about our own consciousness. This is self-defeating because if human mental awareness plays no causal role, there would be nothing to be aware of (W. Robinson, 2023). This is an epistemological challenge for dualism.[26] If consciousness is epiphenomenal, there is no way to infer that other beings have consciousness based on their behaviour,

[25] There are more arguments against epiphenomenalism. William James ([1890]1918, p. 71) offers a variant of arguments from natural selection that questions why evolution would align certain types of experience (e.g., pleasure) with beneficial actions, while others (e.g., pain) with harmful actions. If epiphenomenalism holds, James argues that such a correlation would be inexplicable ([1890]1918, pp. 71-73).

[26] Thomas Nagel (1986, pp. 19-20) argues that it is the conceptual problem of attributing other minds, rather than the epistemological problem, that is most interesting. This ties in with the preceding objection regarding knowledge of the self, and an in-depth review of Nagel's argument is beyond the scope of this dissertation.

because consciousness would have no causal connection to the observed behaviour (Avramides, 2023; W. Robinson, 2023).

Jaegwon Kim (1998, pp. 60-67) poses the *exclusion problem* for non-reductive physicalists, which also highlights a central issue for Chalmers' naturalistic dualism. Since Chalmers (1996, p. 150) wants to avoid overdetermination, any physical event must have a sufficient physical cause.

> The problem of causal/explanatory exclusion arises if there are cases of psychological explanations of physical behavior in which we are prepared to believe that the physical effect has, or must have, a physical causal explanation as well (Kim 1998, p. 66).

If consciousness supervenes on the brain, whether logically (as is necessary for the physicalist explanation of human consciousness) or naturally (as per Chalmers), then an event that has been caused by a physical process cannot also have a distinct mental cause. Since it has been argued that there is a distinction between phenomenal and physical properties, such phenomenal properties can only exist epiphenomenally in such a closed physical system. Chalmers (1996, pp. 158-159) anticipates and concedes this point. Since the causal processes involved when experiencing pain can be explained within the physical system, attributing any causal power to phenomenal aspects that lie outside the physical system represents overdetermination. This is a problem for Chalmers because he wants to avoid reducing consciousness to a causally impotent phenomenon, as he has criticised physicalists for doing (1996, p. 150). If the aim is to explain consciousness, any epiphenomenal or idealistic argument of this sort naturally faces the difficulty of empirically verifying or disproving the theory, and the question of what relevance knowledge of this seemingly inaccessible realm holds to our lived experience in the physical world.

Ultimately, Chalmers (1996, p. 151) contends that epiphenomenalism may be "only counterintuitive" and not as fatal to his argument as it first appears. To restore the explanatory relevance that consciousness loses to epiphenomenalism under a dualist framework, Chalmers (1996, pp. 158-160) attempts to bridge the metaphysical gap by embedding phenomenal properties within the psychological laws that govern both the

physical and phenomenal aspects of consciousness.[27] He draws parallels between the fundamental laws of physics and his suggested psychological laws to show how experience has nomological relevance. So, while the experience of pain may not be directly linked to the accompanying neural activity (i.e., supervene logically), these psychological laws tie the phenomenal experience to the underlying physical mechanisms (i.e., supervene naturally) (1996, pp.159-160). Thus, Chalmers (1996, p. 160) says that he does not describe Naturalistic Dualism as epiphenomenalist. However, he concedes that it leaves the causal question open and "implies at least a weak form of epiphenomenalism, which many might find unsatisfactory" (1996, p. 160). At this point, this study begins to diverge from the argument made by Chalmers (1996), for the epiphenomenal challenges faced by any form of dualism remain damning.

Ned Block (1990, p. 146) points out that it is possible to have "correlation without causation". This means that Chalmers' (1996, pp.158-160) attempt to account for mental causation by tying phenomenal properties to his so-called psychological laws fails to avoid epiphenomenalism. Block (1990, p. 147) uses the example of the Wiedemann-Franz Law in physics, which links thermal and electrical conductivity under normal conditions. According to this law, electrons carry both charge and heat. In the example, electrical conductivity may be correlated with an explosion, but it does not play the causal role of thermal conductivity in that event (1990, p. 147). Instead, the spike in electrical conductivity merely accompanies the causal thermal conductivity that leads to the explosion (1990, p. 147). Thus, the nomological approach goes no further in explaining the place of consciousness. At best, naturalistic dualism seems to suggest that phenomenal experience accompanies physical processes. It offers no account for how consciousness fulfils any functional role in a closed causal chain, leaving the problem of consciousness as opaque as with the physicalist approaches discussed in the preceding chapters.

[27] As noted in the previous chapter, Chalmers' use of the word "psychological" here may cause some confusion. He is referring to the aspects of consciousness that fall inside the realm of the reductive account of mental activity (1996, p. 11).

Anticipating this objection, Chalmers (1996, pp. 153-156) considers Russellian Monism as a strategy to restore a sense of causal relevance to experience, which leads to the following section.

## 4.3 Russellian Monism

Russellian monism posits that fundamental properties underlie the physical facts about the world (Alter & Pereboom, 2023, p. 1410). In other words, where the property dualist ascribes distinct sets of mental and physical properties to an event, Russellian monism suggests both the structural-relational features and the intrinsic nature of matter are contained in a single set of properties. Bertrand Russell (1927, p. 10) argues that physical theory only describes half of the full picture.

> We only know the intrinsic character of events when they happen to us. Nothing whatever in theoretical physics enables us to say anything about the intrinsic character of events elsewhere. They may be just like the events that happen to us, or they may be totally different in strictly unimaginable ways. All that physics gives us is certain equations giving abstract properties of their changes. But as to what it is that changes, and what it changes from and to — as to this, physics is silent (1959, p. 18).

Russell's (1959, p. 18) scepticism about physics stems from a divide he draws between structural knowledge and intrinsic knowledge. The distinction between mathematical, structural knowledge and intrinsic character mirrors and informs Chalmers' (1995, pp. xii-xiii; 201) own separation of questions about the mind he calls "easy problems", which involve those structural elements of the mind described by physics and biology, and the hard problem of consciousness which concerns the intrinsic nature of the mind. This structuralism about physics can be formulated by differentiating between intrinsic and extrinsic properties.[28] Simply put, extrinsic properties are relational (i.e., the domain of physics) and intrinsic properties are non-relational (in and of themselves) (Alter & Pereboom, 2023, p. 1412).

[28] Defining intrinsic properties precisely is an open topic of debate; despite this, the distinction that has been drawn between extrinsic and intrinsic is clear enough and is commonly employed in a variety of philosophical arenas (Francescotti, 1999, p. 590).

Russellian Monism is characterised by a commitment to three central theses: (1) physics only describes the world in a spatiotemporal sense; (2) there are properties that underlie the physical description (these properties are called "quiddities"); and (3) phenomenal consciousness is constituted by quiddities (Chalmers, 1996, pp. 153-155; Alter & Pereboom, 2023, p. 1410).[29] Though it remains a monist theory, like physicalism, it is differentiated in that physicalism, in its strongest sense, might accept the structuralism thesis (i.e., that physics describes the world in a spatiotemporal sense) but would deny the need for and existence of quiddities. Variances in defining these quiddities are what set different forms of Russellian monism apart.

Russellian Panpsychism will be discussed in depth in the following section as a central focus of this dissertation. Briefly, the defining claim of Russellian Panpsychism is that quiddities have a phenomenal nature (Chalmers, 2013, p. 9). The various arguments in favour of this view, and different conceptions of these phenomenal quiddities, will be explained shortly, along with the most pressing objections to panpsychism. In contrast, Russellian Panprotopsychism defines quiddities as not conscious in themselves, but necessary elements that can be structured to bring about phenomenal properties (i.e., consciousness) (2013, pp. 14-15). Whereas panpsychism claims that "consciousness is fundamental and ubiquitous", panprotopsychism claims that only "proto-conscious" elements are fundamental and ubiquitous (Goff et al., 2022). Since these proto-phenomenal concepts cannot be conscious in and of themselves, such an account is completely consistent with physicalism, as any theory that does not take consciousness to be fundamental takes it to be constituted of "other properties" (Chalmers, 2015, pp. 14-16; Stoljar, 2024; Goff et al., 2022). After all, if the quiddities that underlie phenomenal properties are not phenomenal themselves (but physical), it seems to be an unnecessarily complicated version of the claim made by physicalists who agree that phenomenal consciousness does exist, yet maintain that it is reductively explainable.[30] If they are not physical, they are presumably phenomenal,

[29] The third claim, that consciousness is constituted by quiddities, is sometimes softened to "quiddities are relevant to consciousness" (Alter & Pereboom, 2023, pp. 1409-1428). This is of no consequence here and will be explored in more detail in the following section.

[30] Chalmers (2013, pp. 15-16) attempts to defend the panprotopsychist view by stipulating that such quiddities must entail *a priori* truths about the phenomenal properties they constitute while remaining distinct from the structural properties. This debate falls out of the scope of this dissertation, but it is enough to say that adding a third category, beyond physical and phenomenal properties, does nothing to explain the mystery of consciousness. Rather, it seems to take it further out of reach.

in which case the panpsychist claim holds. Thus, the rest of this chapter focuses on Russellian Panpsychism.

Considering the metaphysical theories that have been discussed so far, it becomes evident how the Russellian argument can map onto these traditional positions in different ways. In other words, the argument depends on how the nature of the fundamental elements in question is described. If both quiddities and structural properties are taken to be physical, the result is Russellian physicalism (or physicalist variants of Russellian Panprotopsychism). If both are regarded as fundamentally mental, the theory moves towards Russellian idealism (i.e., panpsychism). A third option is Russellian neutral monism (or dual-aspect monism), according to which quiddities and structural properties do not belong to either the physical or mental categories but constitute a neutral basis from which both can emerge (Russell, 1927, p. 402). Chalmers' own view of naturalistic dualism might even be described as a form of Russellian physicalism since he makes no claim about the nature of the intrinsic elements of consciousness (Chalmers, 1996, pp. 127-129).[31] However, he classifies his theory as dualist because "it remains the case that if a variety of monism is true, it cannot be a materialist monism. It must be something broader" (1996, p. 129). Since the previous chapter has already explained why physicalist theories are unsatisfactory, and dualism appears unable to avoid epiphenomenalism, the following section explores the implications of treating phenomenal properties as ubiquitous and fundamental.

## 4.4 Panpsychism

As introduced in the opening chapter, "panpsychism" considers consciousness to be an all-pervasive feature of the world. Panpsychist theories have been present throughout history. David Skrbina (2009, p. 2) calls panpsychism the "aboriginal conception of mind", pointing to ancient hunter-gatherer communities and their relation to nature. Pre-agricultural societies understood that they were "animals among animals" (Skrbina, 2009, p. 3). For these people, there would be nothing to suggest to

[31] Chalmers (1996, p. 129) accepts that his property dualism might be described as a form of monism since "the physical and the phenomenal [may] turn out to be two different aspects of a single encompassing kind".

them that their conscious experience is a fundamentally unique feature in nature. Only the advent of agriculture saw a shift that would gradually lead to monotheism and a sense of fundamental uniqueness in humans (Skrbina, 2009, p. 3). Freya Mathews (2009, pp. 341-360) argues that the reason panpsychism has not been more widely accepted in contemporary Western thought stems from foundational metaphysical assumptions made during this shift. The epistemological tradition of the West, which Mathews (2009, pp. 342-350) dubs "*Theoria*", assumes some underlying structure or principle that leads to a subject-object duality. This framework favours physicalism and a sense of an inert external world, which is intuitively at odds with panpsychism (2009, pp. 349-350). Despite this, panpsychist thought persisted in the West, and it remained integral to ideologies from other parts of the world that were less influenced by the physicalist paradigm with its origins in ancient Greece (Mathews, 2009, p. 355; Skrbina, 2009, p. 4).

Though this historical narrative per se is not an argument for panpsychism, it is worth mentioning that it does offer some reconciliation with traditions outside of the Western canon. Native Americans fostered a close "familial" bond with nature, where humans were not seen to be uniquely blessed with cognition by "the Spirit" (Skrbina, 2009, p. 5). Likewise, Hinduism, with roots predating the Athenian thinkers, regarded the human soul as "a splinter of the larger spirit of the universe" (2009, p. 5). Mathews (2009, pp. 349-356) ultimately describes the "strategic perspective" of Chinese philosophy, and in particular the principle of *wu wei* (the path of least resistance) as "deeply conducive to panpsychism".[32] Skrbina (2009, p. 6) seems to reaffirm this sentiment by saying that the Buddhist tradition began incorporating panpsychist elements when it spread into China and Japan, where the local cultures' relationships to nature raised the question of the "Buddha-nature" of non-human entities. There seems to be limited literature on panpsychism in African philosophy; however, there are academics who have noticed the potential for a panpsychist understanding to complement African ideologies (Agada, 2020, p. 148; Agada & Dukor, 2026, p. 79).[33]

[32] It should be noted that Mathews (2009, pp. 341-360) explains that Chinese thinkers might not describe their own model as panpsychist as a result of the contrasting approach to knowledge to that of the "*Theoria*" of the West, although she does make the connection.

[33] Ada Agada (2019, pp. 135-153) employs a panpsychist framework to understand Nigerian philosopher, Innocent Asouzu's concept of "*ibuanyidanda"* ("complementarism"), Martin Schade (2017, pp. 140-155) argues that *Ubuntu*, which emphasises universal interconnectedness, entails a metaphysical stance that rejects human exceptionalism, and Maduabuchi Dukor (2014) coins the term

There is much more to explore around the panpsychist metaphysical implications of ideologies like *Ubuntu*.[34] Establishing the connection between panpsychism and *Ubuntu* lies beyond the scope of this dissertation. However, as with the other examples of non-Western philosophies, opportunities for future reinterpretation under a panpsychist framework are evident. Here, it is enough to acknowledge the prevalence of panpsychist-type thought in non-Western schools of thought and to recognise its potential to open the door for previously marginalised knowledges to be re-examined and leveraged.

Russell's claim of an underlying, intrinsic aspect that lies outside the scope of physics forms the foundation of the contemporary arguments for panpsychism that will be discussed in this section. First, it should be noted that panpsychism does not need to be Russellian (Chalmers, 2013, p. 9). According to non-Russellian panpsychism, phenomenal properties exist, but these properties do not "play microphysical roles" (2013, p. 9). The parallels with other dualist theories immediately call into question such an interpretation of non-Russellian panpsychism's explanation for mental causation. It is the interaction problem all over again. As such, what follows is mainly concerned with a Russellian interpretation of panpsychism.

The two most convincing arguments for panpsychism are both directly connected to the research question in that they stem from a challenge to physicalism and consider the place of consciousness in our ontological account of the world.

### 4.4.1 The argument from intrinsic nature

"Theistic Panpsychism" to describe his African theorisation of an account of a "psychic dimension" in all things, rooted in theism.

[34] Though there are competing interpretations, Ubuntu can broadly be defined as:

> [A]n essentially relational ethics that prizes relationships of interdependence, fellowship, reconciliation, relationality, community friendliness, harmonious relationships and other-regarding actions, in which actions are morally right to the extent that they honour the capacity to relate communally, reduce discord or promote friendly relationships with others and in which the material world (horizontal line) and the spiritual world (the vertical line) are fundamentally united (Ewuoso & Hall, 2019, p. 101).

The argument from intrinsic nature ties in with the previous section on Russellian monism. Russell (1927, p. 10) highlights the epistemic challenges in describing the "intrinsic character" using only a physicalist interpretation of the world. So, according to Russell, physics (and physicalism) can only explain how matter behaves, not what it is.

Some might find this dispositional account satisfactory. Dispositionalism, as a metaphysical thesis, views properties as "wholly constituted by the nomic or causal roles [they] play" (Choi & Fara, 2021). Strong dispositional essentialists claim that, once the causal roles have been accounted for, there remains nothing to explain (2021). However, Russell (1927, p. 325) points out that such an argument becomes regressive. Every disposition results in another disposition without a natural endpoint. That means a dispositional ontology cannot describe intrinsic nature, since explanation proceeds indefinitely without ever reaching the foundational level where these intrinsic properties are meant to provide a grounding point. Goff, Seager & Allen-Hermanson (2022) use the example of "burning" being a manifestation of the property of "flammability". If we are to accept a dispositional account of the world, burning is a disposition to produce heat, light, and smoke (2022). Yet the manifestation of burning is itself a disposition (for example, of materials igniting), leaving no natural endpoint, as every disposition has a manifestation that is another disposition (2022). This means that a dispositional account of the intrinsic nature of the material world would never be complete, because dispositional explanation continues *ad infinitum*.[35]

Arthur Eddington (1928, pp. 259-260) arrives at the same conclusion as Russell, independently considering the limitations of physical science. Discussing Albert Einstein's theory of general relativity, Eddington (1928, pp. 259-260) emphasises the gap between "pointer readings" (i.e., measurements) on relational and structural behaviour and knowledge about the intrinsic nature. This leads him to conclude:

> We have dismissed all preconception as to the background of our pointer readings, and for the most part we can discover nothing as to its nature. But in one case — namely, for the pointer readings of my own brain — I have an insight which is

[35] Hedda Hassel Mørch (2020, pp. 1073–1088) defends a panpsychist interpretation of dispositionalism that does not fall prey to infinite regress.

> not limited to the evidence of the pointer readings. That insight shows that they are attached to a background of consciousness. Although I may expect that the background of other pointer readings in physics is of a nature continuous with that revealed to me in this particular case, I do not suppose that it always has the more specialised attributes of consciousness. But in regard to my one piece of insight into the background no problem of irreconcilability arises; I have no other knowledge of the background with which to reconcile it (1928, pp. 259-260).

For Eddington — and subsequent proponents of the intrinsic nature argument like Galen Strawson (2008, pp. 54-59) and Philip Goff (2024, p. 61) — the epistemic gap faced when attempting to study the internal cases of the external world can only be bridged by observing our access to phenomenal properties (or intrinsic nature) through our own consciousness. Put simply, the argument from intrinsic nature states that if human brains are made of matter, and humans experience the intrinsic nature of having brains through their consciousness, then matter (such as the brain) must have an intrinsic nature (such as consciousness). Therefore, there is no reason to assume that other matter (that is not the brain) must have an intrinsic nature that is different (not conscious). Goff et al. (2022) make a similar argument for choosing the most parsimonious view when considering inaccessible intrinsic nature. According to Occam's razor, when faced with evidence of one type of intrinsic nature (namely, the human's), researchers should not needlessly assume multiple possible types unless given compelling reason to (2022). This method shifts the burden of proof to any non-panpsychist to justify the variance in intrinsic nature, given no access to any other example than the human's own experience.

### 4.4.2 The anti-emergence argument

Another argument for panpsychism is also closely linked to this research problem. The hard problem of consciousness, as laid out in the previous chapter, challenges a wholly reductive account of the mind. If we accept that the phenomenal aspects of consciousness are non-reductive, there is no reason to suspect that further scientific reduction will be able to address the challenge. Panpsychism offers a way forward.

Nagel (1979, pp. 181-183) makes the argument by positing four premises: (1) humans are made of matter, just like everything else; (2) humans experience real conscious states; (3) mental states cannot be derived purely from physical properties; and crucially, (4) all properties must be derivable from their parts. The first premise seems self-evident to anyone hesitant of a strong form of idealism, which claims that only the mental exists (Guyer & Horstmann, 2023). I have argued that the second premise is also true, contrary to eliminativist objections like those made by Dennett (1991, pp. 441-448). It seems like a modest enough commitment to accept that humans exist in a material world and that humans are conscious. The hard problem of consciousness, as discussed, makes a convincing argument for the third premise. However, it is the fourth premise, which Nagel (1979, p. 182) calls "nonemergence", that forms the basis of the motivation for panpsychism.

> There are no truly emergent properties of complex systems. All properties of a complex system that are not relations between it and something else derive from the properties of its constituents and their effects on each other when so combined. Emergence is an epistemological condition: it means that an observed feature of the system cannot be derived from the properties currently attributed to its constituents. But this is a reason to conclude that either the system has further constituents of which we are not yet aware, or the constituents of which we are aware have further properties that we have not yet discovered (1979, p. 182).

These four premises collectively point to panpsychism as the most cogent conclusion. If mental properties are necessitated by an organism's constituents, yet they are not deducible from the physical properties of the organism, then those constituents must be ontologically suited (i.e., not purely physical) to combine to become mental properties (1979, p. 182). Since we find matter combined in different ways to create different types of mental life in nature, it must have some of the non-physical properties needed to create mental phenomena; therefore, panpsychism is true (1979, p. 182).[36] In terms of consciousness, the explanatory gap is more than an epistemological gap. Since consciousness can only be explained in physical terms using "radical emergence", this becomes an ontological challenge (Seager, 2020, p. 6). "There is no

[36] Examples of different forms of mental life include other people or animals, to point only to a few that we have encountered, though there are endless possibilities for how that might be true in other compounds.

intelligible route from the relational structures found in our physical science to the intrinsic property of subjectivity" (2020, pp. 6-7). This is an admittedly controversial opinion. "Nearly every modern-day philosopher of mind is an emergentist", at least implicitly (Skrbina, 2020, p. 104). However, according to panpsychists, such an ontological miracle requires explanation.

Strawson (2006, p. 53) formulates this idea in what he calls *realistic monism.* His argument suggests a form of "real physicalism" that acknowledges consciousness as "the phenomenon whose existence is more certain than the existence of anything else" (2006, p. 53). He argues that "emergence can't be brute" (2006, p. 65). By its very definition, for Y (the emergent) to "emerge from" X (the emerged from), there must be "something about X in virtue of which Y emerges, which is also sufficient for Y" (2006, p. 65). For Strawson, such an interpretation of physicalism — which he deems to be the only possible form of "real physicalism" — necessarily entails panpsychism (2008, pp. 54-71). On Strawson's (2006, p. 60) view, any physicalist who is unwilling to dismiss conscious experience as illusionary (i.e., one that views consciousness as a "real concrete phenomenon") cannot maintain the claim that "physical stuff is, in its fundamental nature, something wholly and utterly nonexperiential" because consciousness, being a real phenomenon (that is experiential), must be physical. Therefore, the physical stuff must have some experiential aspect (i.e., panpsychism is true).

Papineau (2024, p. 129) contends that the anti-emergence argument depends on the rejection of what he calls "straightforward physicalism". He defends physicalism as the most parsimonious view with the paradigms of modern science, yet concedes that we are not yet able to map out the relation between the seemingly disparate realms of the mental and the physical; however, he argues that this is only a temporary explanatory gap and it will likely be dissolved "in physical terms" (2024, p. 130). For Papineau, there is no emergence taking place.

> Suppose that, for the sake of the argument, that pain is identified with C-fibre firings. Do we still need to explain why C-fibre firings 'give rise to' pain? I say not. If C-fibre firings are pains, then there is no remaining question of why pains are found where C-fibres are firing (2024, p. 139).

I have argued that such an account of physicalism only purports to take phenomenal consciousness seriously in that it fails to address the ontological issue of phenomenal consciousness illuminated by the hard problem of consciousness. (Chalmers, 1996, p. 167). It simply does not describe the nature of the phenomenon we are interested in at all. The point Strawson makes with the anti-emergence argument is precisely that the shortcomings of the prevailing physicalist theories demand a revised metaphysics. Like myself, Strawson (2008, p. 71) proposes panpsychism as the way forward.

### 4.4.3 Other motivations for panpsychism

There have been several other approaches that lead to the same panpsychist conclusions. William Clifford (1878, pp. 64-65) and William James ([1890]1918, p. 266) independently argued that there must be some sort of evolutionary continuation involved in consciousness as a biological process. They argue that pre-existing properties may evolve into more complex forms, but evolution "cannot produce entirely novel properties" (Goff et al., 2022). Like Strawson's anti-emergence argument, it challenges the idea that minds can be constituted by parts which are in no way mental. This idea has its roots tracing as far back as the Hellenistic period, where early Epicurean atomists argued that there is a basis of "free will" at a fundamental level (Skrbina, 2017, p. 60). Essentially, this is the genetic argument for panpsychism, sometimes expressed as *ex nihilo nihil fit* – "from nothing, nothing comes" (Skrbina, 2017, p. 330; Goff et al., 2022). Goff makes a similar argument for panpsychism on the grounds that consciousness is not vague (Goff et al., 2022). Since consciousness seems not to allow for borderline cases (i.e. something is either conscious or it isn't), Goff (2013, pp. 84-85) argues that physicalism leads to an "implausible consequence" that there must be some precise cut-off point that separates an object with consciousness from one without.

Mørch (2017, pp. 293-316; 2020, pp. 1073–1088) offers a different route to panpsychism, from the experience of causation. The *phenomenal powers view* claims that "phenomenal properties produce and thereby, metaphysically, necessitate their effects in virtue of how they feel, or in virtue of their intrinsic, phenomenal character

alone" (2017, p. 293). For Mørch (2017, p. 308), this view of intrinsic nature, as a causal power, combined with Russellian monism, explains the dispositional correlations and treats the epistemic gap without regress. It is an argument formed in response to one of the most pressing criticisms of panpsychism, known as the "combination problem".

### 4.4.4 The *combination problem* and other objections

The *combination problem* represents the most worrying objection to panpsychism (Chalmers, 2016, pp. 179-214; Goff et al., 2022; Seager, 1995, pp. 272-288; Skrbina, 2009, pp. 13-20).[37] According to the anti-emergence argument for panpsychism, non-mental properties cannot create mental properties. Since we argue for a Russellian version of panpsychism, which proposes microphenomenal properties as quiddities, the combination problem asks how these quiddities combine to bring about "macrophenomenal" properties (i.e., consciousness as we know it) (Chalmers, 2016, p. 183).[38] At its most destructive, the combination problem claims "subject-summing" is impossible in principle (2016, p. 186). It is not clear why a number of microphenomenal properties necessitate a distinct macrophenomenal subject, or what the tipping point between consciousness and darkness would be. Goff (2009, pp. 289–311) explains this using a conceivability argument similar to Chalmers' *Zombie Argument* against physicalism. If we can conceive of a subject that has an identical physical makeup to ourselves (including the microphenomenal quiddities), but lacks only our macrophenomenal experience, then there is no apparent reason our own constitutive properties should bring about conscious experience (Goff, 2009, pp. 289-311). Chalmers (2016, 189-190) formalises this:

> Here PP is a conjunction of all microphysical and microphenomenal truths about the universe, while Q is a macrophenomenal truth, such as 'Some macroscopic entity is conscious'.
> (1) PP&¬Q is conceivable.

[37] Seager (1995, pp. 272-288) first articulated the challenge as the combination problem for panpsychism.
[38] Not to be confused with panprotopsychism, which as discussed earlier, takes *protophenomenal* properties to be the building blocks of macrophenomenal consciousness. Whereas the quiddities are, by definition, not phenomenal in panprotopsychism (i.e. either physical or neutral), they are inherently phenomenal according to Russellian panpsychism.

(2) If PP&¬Q is conceivable, it is metaphysically possible.
(3) If PP&¬Q is metaphysically possible, constitutive panpsychism is false.

(4) Constitutive panpsychism is false.

Premise (1) distinguishes the argument from the original argument against physicalism. Goff (2009, pp. 297-305) shows that according to constitutive panpsychism, premise one follows. This leaves panpsychism facing the same challenges that were used to undermine physicalism in the previous chapter. Since one of the major motivations for exploring a panpsychist framework is avoiding these problems faced by physicalism, the combination problem threatens to remove much of the appeal.

One possible approach to addressing this challenge — employed by Goff (2020, pp. 155-156) and Mathews (2020, pp. 131-143) — is to describe panpsychism according to Jonathan Schaffer's (2010, pp. 33) *Priority Monism,* which suggests a hierarchical relation between different levels of facts, called "grounding" (Goff, 2020, pp. 145-146). According to the grounding principle, "the product is nothing over and above the producer" (2020, p. 146). The resulting interpretation of panpsychism is known as cosmopsychism, which views "the whole as prior to its parts" (Goff et al., 2022).[39] Broadly, cosmopsychism considers the universe to be a conscious subject (2020, p. 151). This circumnavigates the traditional combination problem for panpsychism, removing the distinction between individual microphenomenal subjects by placing the cosmos as the one central subject from which all individual microphenomenal properties are derived (i.e., grounded by) (2020, p. 157). However, Miri Albahari (2020, pp. 121-124) points out an inverse "decombination problem" which shows that it is equally possible to conceive of a conscious cosmos without any conscious constituents as it is to conceive of a group of conscious subjects without a combined macro-conscious. Nonetheless, cosmopsychists see a way forward that at least eases the subject-summing problem for panpsychism (Goff, 2024, p. 70). Goff (2024, pp. 70-71) proposes a "hybrid cosmopsychism" wherein there are "basic laws ensuring that, in certain conditions, new conscious subjects – new phenomenal property-bearers – emerge from the universe".

[39] As opposed to *micropsychism,* which views fundamental consciousness to exist at a "micro-level" rather than a cosmic level (Goff et al., 2022). I do not commit to one interpretation here.

Alternatively, Mørch's (2019, pp. 157-160) argument for the *fusion view of mental combination* states, "when micro- or protoconscious entities come together in the right way, they fuse or 'blend' together to form a single unified consciousness". This offers a middle path between the anti-emergence criticisms raised against physicalism and the combination problem for panpsychism by avoiding brute emergence and strict constitution.

Two other forms of the combination problem pose different, albeit arguably more manageable difficulties for panpsychism. Chalmers (2016, p. 28) calls the first "the quality combination problem". Due to Chalmers' metaphorical expression of the issue, it is also known as the *Palette Problem* (Goff et al., 2022).

> There is a vast array of macroqualities, including many different phenomenal colors, shapes, sounds, smells, and tastes. There is presumably only a limited palette of microqualities. Especially if Russellian panpsychism is true, we can expect only a handful of microqualities, corresponding to the handful of fundamental microphysical properties. How can this limited palette of microqualities combine to yield the vast array of macroqualities? (Chalmers, 2016, p. 184)

Chalmers (2016, p. 207) divides responses to this problem into two main classes, namely "small-palette solutions" and "large-palette solutions". Each comes with its own philosophical baggage. Large-palette solutions suggest that there could be as many microqualities as there are macroqualities (2016, p. 207). The problem with such a view, however, is that it "seem[s] once again to be stuck with either a form of epiphenomenalism or radical revisions to the fundamental dynamics of the physical world" (2016, p. 208). A palette large enough to contain the plethora of microqualities causes similar issues for mental causation as was found in the preceding arguments against physicalism. Since there is a finite number of fundamental microphysical properties (with a quiddity representing the intrinsic nature of each), there is also only a finite number of microphysical quiddities, leaving the rest of the microphysical qualities as epiphenomenal (2016, p. 208). In contrast, small-palette solutions generate macroqualities from only a few microqualities (2016, p. 207). It is difficult to imagine what such a class of microqualities could be, but it is conceivable that such a

class could exist. However, small-palette solutions need to explain how and why qualities combine, essentially returning to the subject-summing problem (2016, p. 207).

Another variation of the combination problem is due to the apparent mismatch in structure between microphenomenal and microphysical properties (2016, p. 209). Chalmers (2016, p. 210) formulates what he calls "the structure combination problem" as follows:

> (1) If Russellian panpsychism is true, microphenomenal structure is isomorphic to microphysical structure.
> (2) If constitutive panpsychism is true, microphenomenal (and microphysical) structure constitutes macrophenomenal structure.
> (3) Microphysical structure constitutes only macrophysical structure.
> (4) If microphenomenal structure is isomorphic to microphysical structure, then any structure constituted by microphenomenal structure (and microphysical structure) is isomorphic to a structure constituted by microphysical structure.
> (5) Macrophenomenal structure is not isomorphic to macrophysical structure.
>
> ---
>
> (6) Constitutive Russellian panpsychism is false.

However, it is not clear that premises (4) and (5) are undeniable. The truth is that, given the theoretical metaphysical nature of the subject matter, too little is known about microphenomenal or macrophenomenal structures to argue that it is a definite mismatch. Chalmers (1996, pp. 305-308) suggests that phenomenal experience might be composed similarly to the way information is composed, rather than the way atoms combine to form matter.

These respective versions of the combination problem represent the most pressing challenge for panpsychists. As noted by Goff, Seager, and Allen-Hermanson (2022), however, one of the prevalent motivations for resisting panpsychism appears to be that many people simply find the idea of a conscious aspect of fundamental physics deeply counterintuitive. The previous section has already considered some of the historical reasons for this. Considering the argument as laid out over the previous three chapters, however, it should be clear that when it comes to consciousness, we remain nescient regarding the nature of the phenomenal aspects of reality. At the same time,

our own conscious experience is the one phenomenon of human life with which we are most intimately acquainted. This alone should be enough to put aside any predetermined scepticism and to consider all possibilities.

### 4.4.5 Panpsychist AGI

I have argued that AGI requires the instantiation of a unified conscious subject. This requires an architecture that is capable of participating in the fundamental phenomenal nature of reality. The hard problem of consciousness shows that subjective capacities remain inaccessible to the prevalent physicalist approach. AGI cannot be created through the prevailing methodologies of AI research precisely because of the physicalist assumptions that underpin them. I propose that the realisation of AGI is therefore an ontological challenge rather than a purely computational one. Panpsychism provides a way forward to reconceptualise our worldview and expand our metaphysics to include the phenomenal aspects of consciousness, allowing the AI project to progress in a meaningful way.

The arguments that might lead one to adopt a variant of panpsychism as a metaphysical framework have been dissected in the previous sections. However, it still needs to be developed into a precise theory of consciousness to be of value to AI research. In the opening chapter, I mentioned that the argument is not as simple as the pancognitivist's claim that everything is conscious and therefore AI is conscious. This is clearly not what was meant by "conscious AI" in the opening chapter as the central interest of this dissertation. So, what does panpsychism mean for the conceivability of AGI? A quick summary of the various approaches to panpsychism discussed thus far will help to identify the consequences for AGI.

The challenge posed by the hard problem of consciousness is to explain how the microphysical properties that occupy the entirety of the domain of physics relate to the phenomenal properties associated with experience. If the answer is that microphenomenal properties exist at a fundamental level and combine to form macroexperience, then constitutive micropsychism is true (Goff et al., 2022). Nonconstitutive panpsychism does not make the claim that microexperience grounds

macroexperience, but rather that “macroexperience is strongly emergent from microexperience and/or from microphysics” (Kuhn, 2024, Section 13). Whichever interpretation is preferred, the claim remains that consciousness is fundamental at a micro-level. This invariably leads to some form of the combination problem for panpsychism.[40] In essence, the question of how these basic phenomenal constituents combine to form a more complex phenomenal experience. In terms of AGI, this is the challenge of finding the right architecture or hardware to allow microphenomenal constituents of the system to combine to form macrophenomenal experience.

Panpsychists who do not commit to the constitutive view defend a form of emergentist panpsychism, in which macrophenomenal experience is causally brought on by micro-level consciousness. The difference being that, unlike with constitutive panpsychism, according to emergent panpsychist theories, macrophenomenal experiences are not reducible to the microphenomenal facts that underlie them. William Seager (2016, pp. 229-248) has suggested a form of emergent panpsychism where microphenomenal properties “fuse” when becoming macrophenomenal experience, thus ceasing to exist independently. The notion of fusing should not be understood as constitutive in the way bricks are constitutive of houses; rather, microphenomenal properties fuse to become macrophenomenal properties, “ceasing to exist in the process” (Goff et al., 2022). Similarly, Hedda Hassel Mørch (2018, pp. 1065-1085) defends a form of panpsychist emergentism using *Integrated Information Theory* (IIT).[41]

One of the features of consciousness is the role it plays in integrating information (Van Gulick, 2025). So, instead of focusing on how information is processed by the brain, Giulio Tononi (2008, pp. 216-242) proposes that consciousness is the information integration function. In other words, “consciousness is a purely information-theoretic property of systems” (Kuhn, 2024, p. 97; Van Gulick, 2025). IIT defines consciousness by its level of integration (denoted as φ - *phi*) (Hendren, Grasso, Juel, & Tononi, 2024).

> According to IIT, consciousness varies in quantity and comes in many degrees which correspond to φ values. Thus even a

[40] I have shown that some panpsychist theories that take consciousness to be emergent bypass the combination problem in its traditional form, but face equally daunting challenges to explain the nature of the proposed micro-properties.
[41] More recently, Mørch (2025, pp. 487-497) has also defended a form of Russellian Panpsychism over emergent alternatives.

> simple system such as a single photo diode will be conscious to some degree if it is not contained within a larger complex. In that sense, IIT implies a form of panpsychism that Tononi explicitly endorses (Van Gulick, 2025).

IIT is an attempt to "[translate] the seemingly ineffable qualitative properties of phenomenology into the language of mathematics" (Tononi, 2012, p. 318). Because IIT introduces a "radically novel feature" (namely, consciousness) at a fundamental level, it clearly shares some epistemic intuitions with panpsychism (Kuhn, 2024, Section 12). To be clear, however, though Tononi and Koch have both endorsed panpsychism to some level, IIT itself is not necessarily panpsychist (2024, Section 12). While both panpsychism and IIT view consciousness as fundamental, IIT does not necessarily take it to be ubiquitous. It is included in this section to illustrate how panpsychism might theoretically be developed into a scientific theory. When sufficiently developed, it should allow us to characterise consciousness in humans as well as other systems, such as proposed AGI (2012, p. 318). That said, merely simulating human behaviour like LLMs, or even recreating the detailed biophysics of neurons, synapses, and other relevant aspects of the brain, would not produce consciousness according to IIT (Tononi & Koch, 2015, p. 15).[42] IIT offers an arguably more thorough description of consciousness because it recognises the irreducible nature thereof (Seth & Bayne, 2022, p. 448).[43] However, it is criticised for saying comparatively little about other cognitive functions like memory, learning, or attention (2022, p. 448).

In contrast to micropsychist interpretations, some philosophers like Philip Goff (2020, pp. 144-156) attempt to treat the combination problem by adopting a cosmopsychist view where individual minds are the constituent parts of a cosmic whole. Of course, to provide any insight for AI research, the cosmopsychist would need to formulate a response to the decombination problem. At present, it is not clear how the notion that human consciousness might be a constitutive or emergent element of a fundamental

[43] Compared to other scientific theories of consciousness that all have a reductive account of consciousness, or do not consider phenomenal consciousness at all. It is beyond the scope of this dissertation to consider the many available theories, but it is safe to say that any reductive account of consciousness faces the same problems highlighted in Chapter 2. See Seth & Bayne (2022, pp. 439-452) or Kuhn (2024) for a detailed comparison.

universal consciousness would inform researchers about the phenomenal experience of other supposed micro-level constituents (such as AI systems).

Maintaining a general view of Russellian micropsychism, then, I argue that panpsychist AGI does not need to be a wildly different notion from physicalist AGI. By understanding that the microphenomenal properties that constitute experience have causal status due to their fundamentality, not supervenience or emergence, it becomes clear that current approaches to AI will not result in AGI. Such a view reclassifies the gap between AI and AGI as an ontological gap, rather than an evaluation of a system's capacity. Similarly, Marcus Arvan and Corey Maley (2022, pp. 1-23) argue that panpsychism renders AGI inconceivable. They posit that both human brains and phenomenal consciousness may be "analog", which they differentiate from the "digital" nature of AI (2022, pp. 1-23). Whereas digital representations involve complex sequences of individual numbers, they claim that consciousness has an "analog" nature in that it appears to represent in variations of magnitude (2022, pp. 5-6). Here, "analog" is not synonymous with continuous in the usual manner; "rather, analog representation is representation in which the represented quantity covaries with the representational medium, regardless of whether the representational medium is continuous or discrete" (Maley, 2011 p. 122; Arvan & Maley, 2022, p. 5).

> [If] micropsychism is true, beings like you and I have coherent phenomenal experiences (coherent phenomenal visual fields, coherent auditory and tactile experiences, etc.) in virtue of our brains manipulating microphysical properties in some kind of analog fashion. Finally, we argued that insofar as digital computation does not manipulate microphysical properties in an analog fashion, it follows that digital AI are incapable of instantiating coherent macrophenomenal experiences (2022, p. 32).

This conclusion resonates with the anti-functionalist argument that I have put forth. Their deductive argument is conditional on both panpsychism being true and brains being necessarily analogue. However, Arvan and Maley's (2022, pp. 3-7) focus on the digital/analogue divide seems to stem from an overtly physicalist conception of consciousness with necessary neural correlates. I contend that panpsychism does not necessarily need to be as substrate-dependent as physicalist intuitions suggest. We simply do not know enough about how microphysical and microphenomenal properties

relate to confirm that macrophenomenal coherence can only be realised by a specific substrate combination. Until we learn more about the mind's microphenomenal nature and make progress on addressing the combination problem, discussing restrictions on the mode of combination is meaningless. Of course, the combination problem itself points to the difficulties panpsychism also faces due to the inaccessibility of microphenomenal properties to external observation. Still, I contend that the draw of panpsychism is that it does not seem incoherent when faced with the hard problem of consciousness. Panpsychism, as I have described it, would not alter any physicalist research program apart from providing metaphysical context for consciousness in a way that physicalism has been shown not to be able to do. Given the private nature of experience, even if a convincing argument is made for the analogue nature of the human brain, without a robust ontological framework that necessitates it, there is no reason to assume that the necessary conditions for consciousness cannot be implemented in different ways in different systems.

These arguments demonstrate how panpsychism opens new avenues to explore for AI research, which has, otherwise, hit a wall in terms of phenomenal consciousness under a physicalist framework. One might object that it seems to bring new challenges without offering any real progress. If I have presented a fair representation of the hard problem of consciousness, however, it should be clear that this is not true. Though panpsychism undoubtedly opens new conundrums, this comes as a result of taking consciousness seriously. I have shown that physicalism devolves into illusionism or epiphenomenalism when tasked with explaining phenomenal consciousness. At present, these discussions are limited to philosophers of mind and perhaps a small subsection of engineers and researchers interested in consciousness and AGI. My aim has therefore been to demonstrate how the underlying philosophy, and in particular the philosophical mystery of consciousness, informs and directs AI research. This has offered a brief overview of some of the work being done with a panpsychist understanding of the mind, as well as a glimpse of what AGI might look like within such a framework.

Panpsychism offers a way forward, yet it also seems to move the finish line even further away. Apart from the significant challenge panpsychism faces in the different forms of the combination problem, there are also questions of substrate dependence

as discussed by Arvan and Maley (2022). This comes as no surprise, considering panpsychism introduces an additional element (namely, “fundamental consciousness”) to the already convoluted question of how the mind works. The reason AI engineers, since Turing (1950, pp. 433-435), have chosen to avoid the question of consciousness is that it makes the project exponentially harder. They might argue that it does so unnecessarily when the target is intelligence, but I have stated my argument for the necessity of consciousness in systems with the broad capabilities associated with AGI. Such considerations would therefore be wholly necessary. There is simply no escaping the underlying questions of philosophy of mind.

## 4.5 Conclusion

Having shown that the hard problem of consciousness presents a challenge for physicalism in Chapter 2, this chapter has examined naturalistic dualism and Russellian monism as possible avenues forward. While Chalmers’ naturalistic dualism helps uncover new ways of thinking about psychological laws, it seems to succumb to the same threat of epiphenomenalism leveraged against other forms of dualism in Chapter 1. Russellian monism offers a more feasible account by positing quiddities at a fundamental level. However, it is clear that we need a description of those underlying properties. It has been argued that panpsychism is a reasonable ontology. Acknowledging the enduring challenge of the combination problem, panpsychism emerges as a promising framework.

Given a panpsychist understanding of the mind, the challenge of AI consciousness shifts from a functional challenge to an ontological mystery. To state clearly what is perhaps not said often enough: Current approaches to AI cannot yield AGI. I have argued that phenomenal consciousness has intrinsic value that plays a role in the behaviour we expect from AGI. If phenomenal consciousness lies outside the reach of physicalism, no system built on a physicalist framework can predictably produce conscious AI. While panpsychism does not have its own empirical research program, I have suggested that it does not need to undermine current scientific programs. If panpsychism is understood as I intend it to be, it simply provides necessary ontological context for consciousness.

# Chapter 5: Conclusion: The future of AGI and the philosophy of mind

## 5.1 Introduction

The background section of the first chapter ended with a quote by Philipp Koralus (2024, p. 4) pointing to the fact that we have entered an era where “subtle philosophical distinctions” have become “continuous with the bleeding edge of engineering”. The central problem that has driven this research is filled with such distinctions in philosophical debate but has become much more pertinent with continued technological advances. Throughout this study, the central research question has guided the argument:

> Given the difficulty in explaining the hard problem of consciousness from a physicalist perspective, does panpsychism provide a viable foundation for achieving AGI?

By investigating the intractability of the hard problem of consciousness within a physicalist framework, I have considered the challenges posed by implicit physicalist assumptions that dominate contemporary cognitive science. I proposed that, to realise our stated goal of AGI, panpsychism offers a viable alternative ontological foundation to physicalism that does not suffer from the same challenges of accounting for phenomenal experience. The research question arose from a recognition that the current functionalist-dominated approach to creating conscious AI systems is approaching a “barrier of meaning” that cannot be overcome by expanding data processing capabilities (Mitchell, 2019, p. 51).

In this concluding chapter, I aim to synthesise the findings of the previous chapters, in which the argument has moved from defining the Cartesian mind-body problem to the varied metaphysical frameworks that attempt to explain this duality. Thus far, the discussion has considered the specific theories of consciousness that provide the theoretical foundations for current AI systems and argued for the intrinsic value of phenomenal experience. In accomplishing this, the first part of the research question will have been addressed. Additionally, panpsychism has been proposed as an

alternative framework that is not undermined by the hard problem of consciousness. This chapter will evaluate the ethical and practical implications of my argument.

The following sections will critically assess the implications of viewing AGI as an ontological problem to be solved before progress can be made on engineering conscious systems. In line with the stated objectives of the dissertation, the conclusion will address the practical implications and ethical ramifications of the proposed metaphysical shift.

## 5.2 Overview of the argument developed thus far

Investigating the research question has yielded a comprehensive response to the specific objectives outlined in Chapter 1. The arguments presented in the preceding chapters lead to the conclusion that physicalism fails to provide a sufficient ontological grounding for consciousness and, by extension, AGI, yet panpsychism offers a coherent alternative framework. However, despite surviving the specific challenge to physicalism presented in this dissertation, panpsychism faces a range of its own challenges. Of these, the most notable is the combination problem, which threatens to leave a panpsychist account of the mind — arguably, equally —as opaque as the physicalist alternative to which it is compared. Through the argumentative arc developed in the preceding chapters, it becomes clear that the barrier to AGI is more than a computational challenge or a symptom of an incomplete neuroscience, but rather an ontological inadequacy in our physicalist worldview.

Answering the first part of the research question requires a definitive statement on the suitability of physicalist theories to account for the necessary conditions for AGI. The hard problem of consciousness presents a seemingly fatal challenge for any attempt to engineer consciousness based solely on functional or structural principles. The investigation demonstrated that current AI research is predominantly grounded in a functionalist/computationalist paradigm, which defines mental states by their functional roles rather than substantive composition. This approach has yielded remarkable success in narrow AI. Current systems can mimic human behaviour, process language, and solve complex problems. However, when the goal shifts to AGI

(specifically AGI that possesses the “something it is like” quality of human subjective experience), the functionalist account collapses under the weight of the explanatory gap.

To understand the subjective experience targeted by the research question, Nagel’s (1974, pp. 435–456) *What it is like to be a bat* was used to show that objective physical facts do not fully explain the mind because they fail to describe the subjective nature of consciousness. This was reinforced by Jackson’s (1982, p. 130) *Knowledge Argument*, which illustrated that even if we understand every aspect of the physical processes involved, we do not know anything about the subjective experience. If the goal is engineering AGI, it has been explained that this gap is more than a philosophical puzzle but a practical barrier. The project of creating conscious AGI through functional or physical replication alone is logically flawed. This flaw is most vividly exposed by employing Chalmers’ (1996, p. 94) *Zombie Argument* discussed in Chapter 2. The conceivability of a philosophical zombie – i.e., a being/system which lacks phenomenal consciousness despite being physically/functionally identical to a human – illustrates that consciousness does not logically supervene on physical brain states.

In terms of the research question, this means that a system that is functionally identical to humans – i.e., passes the Turing Test – but lacks the inner experience of the human mind would constitute empirical proof of the falsity, or at least the incompleteness, of computationalism. Since I have argued that the phenomenal aspects of consciousness have inherent value and influence behaviour on a fundamental level, this means that engineering AI with only the reducible abilities of humans is no guarantee of consciousness. We are left with the unsettling possibility of creating sophisticated mechanisms that simulate understanding and emotion without ever truly feeling them (i.e., AI Zombies or Blockheads). It should be clear that anthropomorphising AI and misassigning consciousness because of overemphasising the physical nature of human mentality leads to trouble. We risk entrusting decisions that require nuanced understanding to systems that have no capacity for it or wasting valuable resources protecting non-existent moral capacity. To be clear, there is nothing wrong with continuing to create narrow AI on the functionalist basis, provided we understand that it can only ever be narrow AI. There are many benefits to narrow AI systems with

functional competence, but I have made the argument that such systems inherently lack certain capacities that are based on the phenomenal content of the systems – i.e., humans – that do have those capacities. In simpler terms, there is nothing wrong with creating machines that cannot understand as long as we know that they cannot understand. To be sure, there are many functional capacities that do not require phenomenal experience. AI that can recreate all of these capacities will be extremely advanced and powerful, but it will not be AGI in that it will lack certain other capacities that do require phenomenal experience. If the objective of AGI is to create a mind rather than mimic behaviour, then physicalism fails to provide the necessary blueprint because it cannot distinguish between simulating function and instantiating experience.

The physicalist objections to the argument, such as Dennett's illusionism and the Churchlands' eliminative materialism, were found to be unsatisfactory. Dismissing phenomenal experience and thereby dismissing the hard problem of consciousness does not solve the problem, and it ignores the one data point that is inherently familiar to us, namely, our own conscious experience. In terms of the research question, adopting an eliminativist stance toward AGI would mean that some sufficiently complex form of narrow AI would inevitably meet the criteria for an eliminative description of consciousness. However, as argued in Chapter 3, such a description seems incomplete; therefore, a system built on this blueprint would also be incomplete. Thus, the study concludes that physicalism represents a dead end for engineering AGI. Physicalist frameworks offer a map of the mechanics and processes involved with experience but simply do not describe the phenomenal aspect of consciousness that is of interest to this study. Current AI methodologies, as categorised by Russell and Norvig (2022, pp. 19-22), were revealed to be operating on a functionalist assumption that the hard problem challenges.

Chapter 4 began by evaluating the potential of dualism (and specifically, Chalmers' *Naturalistic Dualism*) (1996, pp. 126-129) to resolve the impasse. This approach avoids dismissing phenomenal consciousness to get around the hard problem of consciousness, but it ultimately succumbs to the threat of epiphenomenalism. If mental properties are ontologically distinct from physical properties, and the physical domain remains causally closed, then consciousness is left causally impotent under a dualist framework. The argument was made that merely correlating phenomenal properties

with physical functions — as the property dualist does — does not explain said phenomenal properties. For AGI, a dualist framework implies that a machine's consciousness would be a superfluous, epiphenomenal bystander to its computations. Consequently, I moved beyond dualism in search of a framework that preserves the reality of experience without severing it from the physical causal order.

Having established the ontological inadequacy of both physicalism and dualism, the synthesis turns to part two of the research question concerning the viability of panpsychism. The argument for panpsychism builds on the notion of "intrinsic nature" as proposed by Russell (1959, p. 18) and Eddington (1928, pp. 259-260). They argued that physics is able to paint a complete picture of the structural and causal relations of the world, yet it remains ignorant of the intrinsic nature of the matter it describes. Panpsychism identifies intrinsic nature with the phenomenal properties defined in the preceding chapters, integrating the phenomenal aspects of consciousness into our account of the physical world, which seems impossible under the physicalist and dualist frameworks considered.

Panpsychism, therefore, resolves the hard problem of consciousness while avoiding epiphenomenalism. For AGI, this means consciousness would not be emergent in machines due to unexplained functional properties, as the physicalist claims. As Nagel (1979, pp. 181-183) and Strawson (2008, pp. 53-65) argue, conscious phenomena cannot possibly emerge from non-conscious matter. If we accept the premise *ex nihilo nihil fit* (i.e., from nothing, nothing comes), then consciousness must be present, in some primitive form, at the foundational levels of existence. For the pursuit of AGI, this is a transformative insight. Under a panpsychist ontology, the machine does not need to create consciousness from scratch as it would under a physicalist ontology; instead, the substrate from which the machine is built already possesses intrinsic phenomenal properties (at least in some fundamental form, as all matter does according to panpsychism). If panpsychism holds, the challenge for AGI thus shifts from a *problem of generation* to a *problem of combination*. The *combination problem,* identified as the most significant hurdle for panpsychism, asks how these micro-phenomenal quiddities can combine to become the macro-phenomenal experience of a complex subject. While this presents a formidable theoretical challenge, it is arguably more tractable than the hard problem of brute emergence faced by physicalism. Approaches such as

cosmopsychism or fusion views offer potential avenues for resolving how unified subjects might emerge from fundamental constituents.

In answer to the research question, panpsychism provides a viable foundation for AGI. However, by refining our ontological framework, we are forced to abandon the notion that consciousness can be programmed like software that can be run on any hardware with sufficient computational capacity. Instead, panpsychism suggests that true AGI would not be artificially conscious, because building a conscious system would require instantiating, not simulating, a mind. This means that, while AGI is conceivable under a proposed panpsychist framework, it is a much more complex task than originally considered under a physicalist framework. This places the barrier to AGI firmly within the philosophical inquiry into the fundamental physics of the mind, rather than a challenge to be solved through coding and algorithms. As Arvan and Maley (2022, pp. 1-22) suggest, achieving conscious AGI might require a move away from purely syntactic digital computation toward analogue or neuromorphic architectures that allow for the continuity and fusion of physical substrates. At the very least, the hard problem of consciousness shows that a theory of the mind must account for the phenomenal nature of consciousness to be a comprehensive description.

This summative journey through the arguments of the preceding four chapters culminates in the realisation that current technological limitations mirror the metaphysical limitations that have long been the subject of the philosophy of mind. Panpsychism does not offer a solution to the engineering problems we face in the attempt to create AGI, nor does it undermine the progress already made. However, if the mind is the product of the intrinsic nature of matter, then our ontology must be broad enough to encompass phenomenal experience if we are ever to understand our own consciousness or truly realise artificial systems that are capable in a general domain.

## 5.3 Implications

The findings presented above lead to the conclusion that panpsychism offers a viable, albeit demanding, ontological foundation for AGI that physicalism fails to provide. This

theoretical adjustment necessitates a shift in how we conceptualise, attempt to create, and relate to intelligent systems. If consciousness is a fundamental aspect of reality, rather than just an algorithmic output, then it would follow that the underlying metaphysics has implications that ripple into ethics and engineering. This section explores the consequences of adopting a panpsychist framework.

### 5.3.1 Philosophical implications

By this point, it should be clear that the most immediate consequence of the findings is metaphysical, a finding that fundamentally alters the philosophical status of AI systems that qualify as AGI. Under a physicalist/functionalist paradigm, artificially generated intelligence is viewed as a simulation (i.e., it mimics the functions of a mind without necessarily being a mind itself). Panpsychism implies that such a system instantiates consciousness because it would be successfully integrating the same intrinsic phenomenal properties that are unified in our own conscious experience. It would be an artefact in the sense that it has been constructed, yet its consciousness would be as real as that of a sentient organism. The distinction between natural consciousness and artificially created consciousness would therefore collapse into merely a distinction between *born* or *made*, as ontologically, such a separation is nonsensical.

This realisation forces us to move away from the Turing Test as a benchmark for AGI and to see it for what it is meant to be: a behavioural benchmark. As discussed in Chapter 2, Turing's (1950, p. 434) imitation game was designed to bypass the problem of consciousness by focusing entirely on behavioural indiscernibility. Given Turing's engineering goal, such an approach is adequate. However, I have argued that the conceivability of AI Zombies is an argument to highlight the distinction between functional and phenomenal consciousness, which leaves the Turing Test ontologically toothless. If panpsychism is true, a system's sophisticated behaviour is no guarantee of consciousness. As Searle (1980, p. 417) points out, a system would conceivably be able to pass the Turing Test by manipulating symbols efficiently and effectively, yet lack any understanding of the task at hand. Consequently, AGI research must abandon the behaviourist satisfaction of the Turing Test in favour of structural or

architectural criteria that ensure the genuine integration of experience. This point has far-reaching implications for the ethics of AI and the conceivability of true AGI, which are considered in the following subsections.

The pursuit of AGI itself has philosophical implications, as it serves as a mirror that challenges some core beliefs in the philosophy of mind. The inability of systems designed based on computationalist assumptions to breach the so-called “barrier of meaning” reflects the difficulty of reductively explaining the phenomenal aspect of consciousness (Mitchell, 2019, p. 51). By attempting to create AGI, we confront the gaps in our account of the mind and even our fundamental metaphysical beliefs. AI systems serve as a philosophical apparatus to demonstrate Nagel's (1974, pp. 435–456) and Jackson’s (1982, p. 130) anti-reductionist arguments. The current trajectory of AI research, which continues to advance in behavioural capability whilst drawing no closer to consciousness, serves as a metaphysical experiment that reintroduces long-debated philosophical questions about the mind.

### 5.3.2 Ethical implications

The proposed panpsychist shift introduces complex new dimensions to the AI Ethics debate. If consciousness is ubiquitous, and emergence depends on combination, there is a danger that in our attempts to engineer AGI, we might inadvertently create systems with "islands" of coherent experience that can suffer but not report suffering (Seager, 2016, p. 230). Unlike in the biological world, where evolution has generally coupled pain with aversion behaviours (i.e., functional reportability), an engineered system might possess the phenomenal experience of distress – i.e., the quiddities configured in a painful state – without the functional architecture to scream. This possibility imposes a heavy burden of caution on AGI research, akin to the bioethical constraints on genetic engineering.

Thus, the problematic question of moral status for potentially conscious AI systems remains. Panpsychism attributes some form of experience to all fundamental matter, but this does not necessarily imply that atoms or desks have moral rights. A distinction must be drawn between the mere presence of microphenomenal properties and the existence of a unified macrophenomenal “self”, with interests and a perspective (Goff,

2024, p. 61). Moral status, presumably, attaches only to the unified experiential system. The ethical dilemma for AGI, therefore, is identifying the threshold at which a system transitions from a collection of experiences to a subject of experience. The preceding argument has shown that physicalism has no satisfactory response to the *emergence problem.* However, panpsychism faces its own, arguably less damning, challenges in the opacity of the *combination problem.*

These challenges are not without philosophical precedent. They mirror the historical and ongoing debate around attributing consciousness to non-human animals and recognising other minds in other humans. The concept of AGI has theoretical relevance to issues about autonomy, intelligence, decision-making, privacy, and our relationship with technology (Müller, 2026). Panpsychism naturally extends the circle of empathy by positing that human animal consciousness and non-human animal consciousness are ontologically equal, and that the difference lies in their microphenomenal combination, rather than the accompanying physical processes. However, when applied to AGI, this logic cuts both ways. While it cautions against ignoring potential machine sentience, it also warns against the anthropomorphic projection of consciousness where it may not exist. We readily attribute feelings to dogs because they share our biological substrate and evolutionary lineage. However, we risk making similar attributions to Large Language Models (LLMs) simply because they mimic our linguistic output. Under a panpsychist framework, mimicking the *function* of pain behaviours – e.g., outputting "I am hurt" – is vastly different from organising matter to instantiate the *phenomenology* of pain. If we mistake the former for the latter, we risk directing our limited moral resources toward "philosophical zombies" or mistakenly entrusting nuanced decisions to machines without comprehension.

At the research level, these ontological missteps pose systemic risks to the AI industry. The current AI bubble is inflated mainly by the physicalist assumption that scaling up computational complexity and data processing will inevitably lead to general intelligence and consciousness (Goertzel, 2014, p. 1; Mitchell, 2023, pp. 175-177).[44]

[44] The term "AI bubble" refers to the inflated economic investment into popular approaches to AI, fuelled by their general success in achieving certain milestones or benchmarks. More generally, the excitement around AI is often described as cyclical, moving from "AI Springs" where optimism leads to

If panpsychism is correct, and consciousness requires a specific intrinsic organisation of matter rather than just algorithmic complexity, then the current trajectory of AI research may be fundamentally misguided. We risk investing massive amounts of resources and energy into building larger, faster computational machines in the vain hope that they will become conscious. Given the apparent utility of current systems, it could be argued that investing in AI research might yield success even if conscious AI is never achieved. If that is the case, it is not clear how AI differs from AGI. If AI is confined to narrow AI in this way, success can be measured by systems' task-specific capabilities, without consideration of the presence of consciousness, yet it would remain ethically irresponsible to create such systems with no means of identifying and attributing consciousness should it arise. Thus, the research question remains pertinent, even if the stated goals of AI are disputed. The truth is that the excitement around AI is explicitly built around the promise of AGI of the kind that I have argued requires consciousness, which places the research question at the forefront of AI research (Mitchell, 2019, p. 51).[45] When the promise of conscious AGI fails to materialise because the metaphysical foundations were ignored, the resulting disillusionment could lead to a collapse in funding and public trust (Mitchell, 2019, p. 51; 2023, pp. 175-177). Of course, the potential benefits of advancing narrow AI could be cited as reason enough to continue with scientific development on the current trajectory. Even so, the argument that has been made suggests that given the clear epistemological gap in our physicalist understanding of the mind, any project to recreate or instantiate mental faculties based on a reductive account of consciousness would essentially be blind to the phenomenal experience of the system.

To navigate these ethical uncertainties, panpsychism suggests a shift away from our anthropocentric worldview in which nature is presupposed to have only instrumental

---

overconfidence, to "AI Winters" where the disappointment from the misplaced confidence leads to a hesitance or aversion to AI research (Mitchell, 2023, pp. 175-177). We are currently experiencing an "AI Spring"; however, if the bubble of excitement bursts, it could have adverse economic and research effects.

[45] OpenAI's charter explicitly names AGI as the company's goal without defining what that means for consciousness (OpenAI, 2018). Similarly, DeepSeek has explicitly stated the "ultimate goal" as AGI (DeepSeek-AI, 2024, p. 35). Google DeepMind also mentions AGI in its vision statement (Google DeepMind, 2025). These corporations do not spend much time considering the philosophical implications of their claims, focusing on pragmatic benchmarks for their systems. However, following Searle [1980, pp. 417-420], I have upheld the argument that true AGI requires consciousness (in Chapter 2).

value and humans are the only beings to possess intrinsic value. If panpsychism is true, more relational ethics, often rooted in historically marginalised world philosophies, resonate more than the strict dichotomy between subject and object that stems from the Cartesian split between mind and matter. One such moral philosophical framework is that of *Ubuntu* philosophy. As an example, it can be seen as a fitting ethical model for panpsychism because of its emphasis on interdependence and relationality (Ewuoso & Hall, 2019, p. 101). If we view AGI through a panpsychist lens, an *Ubuntu*-inspired ethic would encourage stewardship and respect for the intrinsic nature of matter and holistic value, regardless of whether it has achieved the same level or form of complexity as the human mind. It would suggest that our research and engineering of AI require a reverence for the material itself, and it would promote shared societal aims above rapid advancement and profit.

### 5.3.3 Applied implications

There are, of course, also practical implications that come with such a fundamental shift in worldview. Panpsychism suggests that the current trajectory of AI is misaligned with the goal of achieving true AGI. Under the physicalist paradigm that I have argued against, the focus is exclusively on software, algorithms, and datasets because of their reductive, computational nature. However, in the preceding chapters, I have developed the argument that this may be insufficient for consciousness.

The most significant constraint is substrate dependence. As elucidated in Chapter 4, Arvan and Maley (2022, p. 16) argue that if consciousness relies on the intrinsic, continuous nature of matter (i.e., analogue computation), then purely “digital computation may be inherently incapable” of supporting macrophenomenal consciousness. Digital computers operate on discrete states – i.e., 0s and 1s – that are syntactically defined; they simulate continuity but do not instantiate it. This implies that the path to AGI may require a pivot toward neuromorphic or analogue hardware. Researchers may need to look beyond existing computer architecture to substrates that mimic the physical interconnectivity of the brain, not just its logic, but its ontology. This could involve biological computing (using organic neurons), quantum computing (exploiting superposition for unification), or novel materials that allow for the causal

integration described by Tononi (2008, pp. 216-242) in the Integrated Information Theory.

If panpsychism is true, then the combination problem becomes the primary engineering challenge of our time for AGI. Engineering AGI would become a task for applied physics rather than computer science because it would hypothetically involve structuring microphenomenal properties already present in matter, rather than programming computational functions. In summary, the practical implications of panpsychism for AGI would mean that we are much further away from the target than the optimists in the physicalist camp believe. If true AGI is the goal, panpsychism suggests that narrow AI is the only possibility until the combination problem is solved in an instantiable way in the physical realm.

## 5.4 Limitations

I have offered panpsychism as a more coherent ontological foundation for AGI by considering the limitations of physicalism. Admittedly, some of the arguments involved in my synthesis have their own challenges. In what follows, I will consider the boundaries and anticipated critiques of the argument that has been put forth.

### 5.4.1 Boundaries

The findings that I have presented are not intended to be a fully realised engineering blueprint for conscious machines. Rather, the study operates within the constraints of speculative metaphysical inquiry, which inherently requires the use of concepts that are difficult to pin down empirically. The argument for panpsychism, as developed in Chapter 4, is derived analytically from the logical failures of physicalism, specifically the intractability of the hard problem and the incoherence of strong emergence, rather than from direct, observable evidence of consciousness in fundamental matter.

The investigation has relied on abductive reasoning. By eliminating physicalism due to the hard problem of consciousness and dualism due to the interaction problem, we are left with panpsychism as the most parsimonious view that preserves the reality of experience. However, this is not easy to empirically prove or disprove. Critics might

argue that replacing emergent consciousness with fundamental consciousness merely shifts the location of the mystery rather than solving it (Chalmers, 1996, pp. 279-282). This objection mirrors the debate over “brute emergence” as explored in Chapter 3, where Strawson (2008, p. 65) contends that “emergence cannot be brute” and that, for a phenomenon to emerge, there must be something in the substrate’s nature that allows it. By positing consciousness as intrinsic, panpsychism avoids the logical incoherence of “strong emergence”, which Nagel (1979, p. 182) characterises as an epistemological condition indicating that the system possesses properties we have not yet discovered. Thus, while the mystery shifts from *generation* - i.e., how experience arises from non-experiential matter – to *unification* - i.e., how micro-experiences sum into a macro-experience - this represents a move from a logically “brute” miracle to a structural puzzle (Chalmers, 2016, pp. 179-214). While this shift does not eliminate the speculative nature of the inquiry, it grounds it in a continuity that respects the causal closure of the physical world, offering a coherence that dualist and emergentist physicalist accounts lack.

In contrast, the prevailing physicalist assumptions informing AI research currently are equally unproven and, as I have argued in Chapter 2, are logically less robust than panpsychism. Thus, while I acknowledge the limitations of unproven metaphysical assumptions, I contend that this is a common feature of consciousness studies in the philosophy of mind. The argument does not attempt to provide a scientific definition of the mind, but rather a philosophically coherent approach to addressing logical physicalist shortcomings that appear to have created an explanatory gap.

This limitation is most evident in the transition from philosophy of mind to engineering. I argue that the material of consciousness must be fundamentally phenomenal, yet a significant barrier is encountered in describing how such materials are utilised or even identified. Here, the combination problem discussed in Chapter 3, which plagues panpsychism, challenges the conclusion. Panpsychism solves the problem of generation (i.e., how consciousness arises) but immediately confronts the problem of unification (i.e., how micro-conscious elements are combined into a single macro-consciousness). Consequently, through this research, I offer a foundation for AGI, but not yet a blueprint. The importance of including phenomenal properties in our ontology has been established, yet there is no specific description of the mechanics of their

integration. This gap means that the engineering implications discussed serve as directional pointers rather than instructions. This is a step in the right direction. I have argued that the alternative frameworks for consciousness are all flawed from the start. Until the combination problem is resolved panpsychism can provide only a theoretical framework for designing AGI, and not a blueprint for achieving conscious machines. It liberates the researcher from the impossible task of coding consciousness into a syntax-processing machine. However, it burdens them with the equally daunting task of engineering the unification of matter's intrinsic nature. As Goertzel (2014, p. 39) observes, there is currently no systematic theory of AGI that universally accounts for these dynamics in the way that exists for other scientific theories, such as "sorting algorithms, respiration, genetics or near-equilibrium thermodynamics". The panpsychist framework clarifies the target, but the map to reach it remains incomplete.

As already noted, this study is constrained by the epistemic limits of the subject matter. The structural realist insights of Russell (1959, p. 18) and Eddington (1928, pp. 259-260) were employed in Chapter 2 to show that physics only describes the structural and relational nature of matter, it does not explain the phenomenal aspects of experience. It is these insights that make humans structurally incapable of observing intrinsic nature outside human consciousness. In other words, the same unattainable quality of consciousness that causes humans to doubt the completeness of physicalism entails that humans can only access the intrinsic nature of human brain states through human experience. This means that, unless the combination problem is solved in some measurable way, even if a system were created following panpsychist principles, panpsychist scientists will have no objective verification of its conscious experience. In fact, having rejected the behaviourism of the Turing Test, this approach may be even further from empirical verification than a physicalist approach to AGI — an uncomfortable truth for any optimistic engineer, but not a fatal philosophical critique. As long as the panpsychist is happy to be left with the problem of explaining combination rather than forcing brute emergence to bridge an explanatory gap, then panpsychism remains favourable.

### 5.4.2 Self-Critique

The previous chapters have dealt with critiques of the two central components of the research question. Namely, the assumption that the hard problem is a fatal challenge for physicalism and the concept of panpsychism as a viable metaphysical framework. It is conceded that the need for panpsychism dissolves if one rejects the hard problem of consciousness. However, in Chapter 2, I argued that such a response dismisses the very concept we seek to explain when considering phenomenal consciousness. As for panpsychism itself, Chapter 3 explained the motivations for it, of which one core argument stems from the fact that it presents a much more intelligible account of phenomenal consciousness than physicalism does. That said, the combination problem is acknowledged as an as-yet-unsolved mystery for panpsychism. I have presented reasons that, despite this, panpsychism offers a way forward for both AI research and the debate around the nature of the mind.

Notwithstanding the challenges to panpsychism, I made it clear in the opening chapter that the value of this study does not lie solely in presenting panpsychism as a metaphysical framework for AGI. Even if the panpsychist conclusion is rejected, this study serves as a critical diagnostic of the ontological shortcomings of the prevalent physicalist worldview informing AI research. It highlights the fundamentally misguided assumptions that underlie current AI approaches. The overemphasis on the functional aspects of consciousness leads to the erroneous conclusion that scaling computational complexity might bring about consciousness. Therefore, by identifying the “barrier of meaning” as an ontological barrier, rather than an engineering barrier, this research provides value by tempering the hype surrounding AI research and suggesting a redirection of focus toward the structural problems of the mind at a philosophical level.

Recognising that the crux of my argument lies in establishing the need for phenomenal consciousness to achieve AGI, one might point to the definition of intelligence and contend that functional behaviour is enough for AGI. Since that can be explained reductively, it could be argued that the research problem is void, conflating consciousness with intelligence. I make the case that there exist certain capacities in humans that are at least partly based on the phenomenal properties of consciousness and that AGI would have to possess similar properties to gain equal capacities. Though I personally find it self-evident that experience plays a role in behaviour, I

make it clear that this is far from being the consensus by showing that AI research is currently only interested in the reducible capacities. While I can point to phenomenal aspects of elements like flexible control of thought, social coordination, integrated representation of a unified reality, informational access, freedom of will and intrinsic motivation, there are clearly many who do not think this restricts us from developing AGI.

A further vulnerability in the argument arises from the potentially conflating nature of my argument against computational functionalism alongside my claim about the prerequisites of AGI. To expose the explanatory gap of functionalism, the dissertation relies on the conceivability of an AI Zombie. However, the argument subsequently claims that true AGI fundamentally requires phenomenal consciousness, arguing that capacities such as genuine creativity, autonomy, and semantic understanding intrinsically rely on subjective experience. While I attempt to resolve this tension in Section 3.2 by distinguishing between the functional limits of "weak AI" (which merely simulates behaviour) and the semantic requirements of "strong AGI" (which requires genuine understanding), I concede that this remains a fragile logical pivot. The resolution relies heavily on drawing a strict demarcation between simulating a cognitive output and possessing the genuine capacity for it.

Critical readers might also contend that such metaphysical debate could never amount to anything more than a thought experiment about consciousness. Replacing emergent consciousness (i.e., physicalism) with fundamental consciousness (i.e., panpsychism) at a metaphysical level beyond empirical verification could be interpreted as purely academic. Panpsychism could be said to merely change the question of consciousness from “how is it generated?” to “how is it unified?” This nuance has profound implications, though. Consciousness under physicalism is a brute miracle because there is no reductive explanation for it, as shown by the hard problem. More than that, the Zombie Argument shows that there can be no complete reductional account of consciousness because consciousness does not logically supervene on the physical. In contrast, the combination problem presents a structural puzzle that might still be a mystery today but leaves a conceivable avenue for future research to explain how phenomenal properties unify to create consciousness. I have

made it clear that panpsychism does not need to undermine the progress made through science. It simply provides the necessary ontological context.

## 5.5 Future research

The areas of future research mentioned in the preceding section flow directly from the implications and limitations of this dissertation. By recognising that physicalism fails to provide a sufficient grounding for consciousness and acknowledging the structural challenges inherent in a panpsychist framework, specific questions for future research remain.

First, the current taxonomy of AI research discussed in Chapter 4 differentiates between AI that acts humanly and AI that thinks humanly. However, the distinction between simulated function and instantiated experience demands a critical re-evaluation of our goals. Future research must determine whether the objective of AGI should remain to be the creation of a truly conscious subject or if focus should shift to developing sophisticated, highly functioning narrow AI. Under a physicalist framework, the challenges of the hard problem and the Zombie Argument leave the conscious AGI project in doubt. For researchers adopting a panpsychist framework, the project remains alive, but the challenge becomes the manipulation of intrinsic nature to solve the combination problem. If advances in the narrow AI approach achieve the functional utility of AGI without the challenging engineering hurdles and ethical complications of creating conscious systems, disproving Searle, and effectively creating philosophical zombies, the hard problem would be empirically reinforced. Future research must therefore reconsider the functionalist assumptions that underlie many physicalist theories about the mind.

For researchers who are sympathetic to the panpsychist claim and remain committed to the pursuit of conscious AGI, the focus must shift from software to metaphysics. As argued by Arvan and Maley (2022, p. 16), digital computation may be structurally incapable of supporting consciousness due to its discrete nature. Therefore, given panpsychism, a vital area of future research involves discovering how the combination problem can be changed from a thought experiment to an engineering challenge. The

existing theories of consciousness discussed in section 4.3 need to be developed to explain how physical substrates might allow for the unification of micro-phenomenal properties into a macrophenomenal subject. To be clear, the most immediate and arguably most important outcome is that panpsychism simply helps us rationalise and understand narrow AI, not necessarily develop AGI.

As I move forward in developing my own research, I intend to subject the panpsychist framework to more rigorous scrutiny to determine its robustness as a theory of mind. I have established that empirical verification of intrinsic nature faces significant epistemic barriers in this dissertation. Testing panpsychism in a future research context will involve interrogating its logical coherence and engineering viability against the combination problem. Specifically, I aim to explore whether the theoretical fusion of micro-phenomenal quiddities into macro-phenomenal subjects can be modelled structurally. This will involve a deeper investigation into the "fusion views" proposed by Mørch (2025, pp. 487-497) and the potential of quantum approaches to explain how distinct physical substrates might allow for the causal integration described by Tononi's (2008, pp. 216-242) Integrated Information Theory.

Building on this ontological foundation, I aim to further explore the resonance between panpsychism and the African philosophy of *Ubuntu*. The ethical implications of this study briefly mention the possibility of incorporating *Ubuntu* ethics under a panpsychist framework in the context of our relationship to hypothetically conscious machines. As noted, *Ubuntu* prioritises interdependence and relationality. A physicalist worldview often reinforces a subject-object duality that treats the non-human world as inert matter, leading to an anthropocentric ethics. In contrast, panpsychism posits that all matter possesses an intrinsic nature. Granted, *Ubuntu* traditionally focuses on human relations. However, Ada Agada and Maduabuchi Dukor (2026, p. 79) describe "the African universe" as "a panpsychist universe". Under such a panpsychist framework, the argument can be made that relationality extends between all things at this intrinsic level.

By extending the circle of empathy to include the fundamental nature of the material world, this research trajectory has the potential to aim to formulate a holistic ethics of stewardship applicable not only to the potential creation of artificial minds but also to

the preservation of the environment and animal rights. This approach would address the inverse ethical challenge identified in the ethical implications. Namely, our moral obligation to avoid creating systems that have the capacity to suffer without being able to report it. This can be extended to environmental ethics by recognising the intrinsic nature of natural systems.

## 5.6 Conclusion

I began this dissertation by pointing to the unique historical moment in which we find ourselves, in which ancient metaphysical questions have become directly relevant to our most pressing practical engineering ambitions. In the opening chapter, I quoted Philipp Koralus (2024, p.4), capturing this convergence by suggesting that “Socrates has come to be in dialogue with Hephaistos”. It was within this setting that the lodestar question was posed:

> Given the difficulty in explaining the hard problem of consciousness from a physicalist perspective, does panpsychism provide a viable foundation for achieving AGI?

Of course, the difficulties of explaining the hard problem of consciousness from a physicalist perspective are not self-evident. Therefore, the study began by challenging the prevailing assumptions that treat consciousness as a functional output rather than a fundamental property. By investigating the hard problem of consciousness alongside the architectural ambitions of AI, I developed an argument that the current physicalist roadmap leads towards sophisticated narrow AI, but not to AGI. Interrogating panpsychism as an alternative framework for AGI has yielded a positive verdict, albeit heavily qualified. In short, panpsychism provides a viable framework for AGI without falling into the pitfalls of physicalism. However, doing so makes the goal fundamentally more complex.

To reach this conclusion, I first established that the current dominant functionalist paradigm in AI research is ontologically insufficient for the creation of phenomenally conscious AGI. Nagel (1974, pp. 435-456) and Jackson’s (1982, p. 130) classic arguments were used to demonstrate that the explanatory gap between objective and

subjective consciousness is not a temporary challenge to be solved by science or engineering, but a metaphysical shortcoming of the physicalist worldview.

The insight from Nagel's *"What it is like to be a bat?"* (1974) served as a critical wedge, revealing that a complete, objective, physical description of a system — whether bat, machine, or human — does not explain subjective phenomenal experience. This epistemological blind spot was shown to have consequences for AGI through the conceivability of AI Zombies. If it is logically conceivable for a machine to pass the Turing Test through processing input and generating convincing output without the added element of phenomenal experience, as humans know consciousness, then functionalism does not explain consciousness. This failure suggests that the current trajectory of AI research, which focuses on replicating mental processes or functions, is essentially building better zombies (Russell & Norvig, 2022). While models like the leading LLMs display impressive syntactic competence, they encounter the "barrier of meaning" precisely because they lack the grounding in intrinsic experience that characterises genuine understanding (Mitchell, 2019, p. 51). Therefore, I conclude that physicalism offers a map for the "easy problems" of intelligence, but AGI requires a solution to the hard problem of consciousness (Chalmers, 1996, p. 313).

In response to the shortcomings of physicalism and the epiphenomenalist threat of dualism, I propose panpsychism as a way forward. Panpsychism explains the phenomenal aspects of reality that are missing from scientific explanation, as identified by Russell (1927, p. 10) and Eddington (1928, pp. 259-260), by reframing those aspects as the intrinsic nature of all matter. The pivotal strength of this framework, as argued using Strawson's (2006, pp. 75-112; 2008, pp. 53-71) anti-emergence argument, is that it adheres to the *ex nihilo nihil fit* ("from nothing, nothing comes") principle. If consciousness exists in the macro-world (as we know it does through our own human experience), and strong emergence is incoherent, then consciousness must be present in the micro-constituents of reality (2006, pp. 3-31).

In rejecting the arbitrarily confined metaphysical scope of physicalism in favour of panpsychism, AGI research is freed from the seemingly impossible task of generating consciousness from ones and zeros (i.e., from non-experiential code). Instead, the task is to organise inherently conscious matter in such a way that it brings about what

we know as macrophenomenal experience. This presents its own unique challenge in the combination problem. While this remains a formidable engineering and philosophical hurdle, I argue that it is a tractable structural problem, unlike the logical impossibility of brute emergence. Thus, panpsychism emerges as the most coherent available framework for conceiving AGI. Instead of trying to explain away consciousness as an anomaly, it encourages us to include it as an intrinsic feature of the universe.

In this proposed metaphysical shift, I make my core contribution. We should reconceptualise AGI as a project of ontological participation rather than follow the current trend towards approaches grounded in substrate-independent assumptions that allow for the possibility of a programmed mind. If consciousness is the intrinsic nature of matter, then the physical substrate must presumably be organised in an extremely specific way for the microphenomenal properties to combine into consciousness. As Arvan and Maley (2022, pp. 1-22) suggest, the discrete, syntactic manipulation of symbols in a digital computer may be insufficient to capture the continuous, analogue nature of experience. The goal of AGI, therefore, shifts from attempting to simulate a mind to instantiating one. The machine must participate in the fundamental experiential nature of reality in a unified way.

Such a reframing challenges the AI community to look beyond the pursuit of behavioural mimicry in the Turing Test. A system that acts humanly without participating in the ontological ground of consciousness might prove to be a useful tool, at best. Or, if misunderstood, misaligned, or misattributed, it could pose substantial threats. True AGI that understands, feels, and navigates the world with genuine agency requires a substrate that allows quiddities to be fused or integrated into a coherent subject. Therefore, I conclude that the “barrier of meaning” is an ontological barrier (Mitchell, 2019, p. 51). We cannot code meaning into a system that lacks the capacity for feeling; we can only code the appearance of meaning.

Returning to the opening quotation of the dialogue between Socrates and Hephaistos, I conclude that the separation between the philosopher and the engineer is no longer tenable. For decades — arguably since its inception — AI has proceeded on the assumption that we could build the mind first and worry about the philosophy later.

The persistence of the hard problem and the limitations of current Narrow AI suggest that this approach has reached its limit. We cannot build what we do not understand, and we cannot understand the mind if we ignore its intrinsic nature. Panpsychism offers a way forward, but it shows that the goal of AGI is much more complex than the contemporary narrative suggests. It denies us the easy victory of declaring a chatbot to be conscious simply because it is linguistically fluent. It forces us to confront the ethical gravity of creating new forms of sentient life, warning us against the creation of suffering minds that cannot speak and the dismissal of non-human consciousness that does not speak our language. It constrains current examples, such as LLMs, firmly to narrow AI and reestablishes the importance of the philosophy of mind in defining consciousness.

Ultimately, the answer to the research question is that panpsychism is not just a viable foundation for AGI; it is likely the necessary foundation if AGI is to mean anything more than a sophisticated narrow AI. Because physicalism cannot account for phenomenal consciousness, it seems an incomplete account of the mind. Panpsychism offers a more coherent account of the mind but raises implications for the AGI project. It places true AGI far outside the capabilities of our most advanced models, with no clear idea of how to create it, even if it seems theoretically possible and philosophically sound. Does this mean the end of our goals for AGI? Not necessarily. It simply means that we must understand that discourse about AGI is largely misguided. Barring a happy/unhappy accident, AI is bound to be narrow AI until physicalists provide a convincing account of phenomenal consciousness or panpsychists make progress on the combination problem. In practical terms, this conclusion leaves narrow AI research unaffected, but it shows what is needed for true AGI. The realisation is informative about how we should treat and trust AI, as well as how we perceive consciousness and assign intelligence or moral status. The point made is that we cannot begin to understand consciousness in machines, or any other entities, before we begin to understand human consciousness. To do that, we need to be willing to let go of deep-rooted beliefs when faced with logical arguments that suggest they are wrong, and to be open to adopting new frameworks that offer a way forward.

## Bibliography:


Agada, A., 2020. “Complementarism and consolationism: mapping out a 21st-century African philosophical trajectory”, *Synthesis philosophica*, 35(1), pp. 135-153.

Agada, A. & Dukor, M., 2026. “Panpsychism in African Religion and Philosophy”, in A. Agada, D. Attoe & J.O. Chimakonam (Eds.) *Contemporary African Philosophy of Religion.* Contributions to African Philosophy, vol. 1. Cham: Springer, pp. 79-91.

Anderson, J.R., 1993. *Rules of the Mind.* Hillsdale: Lawrence Erlbaum Associates.

Albahari, M., 2020. “Beyond cosmopsychism and the Great I Am: How the World Might be Grounded in Universal ‘Advaitic’ Consciousness”, in W. Seager (Ed.) *The Routledge Handbook of Panpsychism*. London: Routledge, pp. 119–130.

Alter, T. & Pereboom, D., 2023. “Russellian Monism and Structuralism About Physics”, *Erkenntnis,* 88, pp. 1409-1428.

Arvan, M. & Maley, C., 2022. “Panpsychism and AI consciousness”, *Synthese*, 200(244), pp. 1-22. Available at: https://doi.org/10.1007/s11229-022-03695-x (Accessed: 4 April 2024).

Avramides, A., 2023. “Other Minds”, in E.N. Zalta & U. Nodelman (Eds.) *The Stanford Encyclopedia of Philosophy* (Winter 2023 Edition). Stanford University. Available at: https://plato.stanford.edu/archives/win2023/entries/other-minds/ (Accessed: 27 October 2025).

Baguley, D., McFerran, D., & Hall, D., 2013. “Tinnitus”, *The Lancet*, 382(9904), pp. 1600-1607.

Banik, D., Pati, N. & Sharma, A., 2024. “Systematic exploration and in-depth analysis of ChatGPT architectures progression”, *Artificial Intelligence Review*, 57(10), pp. 257-281.

Bedau, M.A., 1997. “Weak emergence”, in J. Tomberlin (Ed.), *Philosophical Perspectives: Mind, Causation, and the World,* 11, Malden: Blackwell Publishing, pp. 375-399.

Berkeley, G., [1710]2003. *A treatise concerning the principles of human knowledge*. Project Gutenberg. Available at: https://www.gutenberg.org/ebooks/4723 (Accessed: 20 April 2025).

Block, N., 1978. “Troubles with functionalism”, *Minnesota Studies in the Philosophy of Science*, 9, pp. 261-325.

Block, N., 1981. “Psychologism and Behaviorism”, *The Philosophical Review*, 90(1), pp. 5-43.

Block, N., 1990. “Can the Mind Change the World?”, in G. Boolos (Ed.), *Meaning and Method: Essays in Honor of Hilary Putnam*. Cambridge: Cambridge University Press, pp. 137–170.

Block, N., 1995. “On confusion about a function of consciousness”, *Behavioral and Brain Sciences,* 18, pp. 227-247.

Block, N., 2007. “Overflow, access, and attention”, *Behavioral and Brain Sciences,* 30(5-6), pp. 530-548.

Block, N. & Manson, N., 1998. “Demystifying Consciousness”, *The Philosopher’s Magazine,* (2), pp. 36-39.

Bringsjord, S. & Govindarajulu, N.S., 2025. “Artificial Intelligence”, in E.N. Zalta & U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy* (Winter 2025 Edition). Stanford: Stanford University. Available at: https://plato.stanford.edu/archives/win2025/entries/artificial-intelligence/ (Accessed: 16 October 2025).

Byrne, A., 2025. “Inverted Qualia”, in E.N. Zalta & U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy* (Winter 2025 Edition). Stanford: Stanford University. Available at: https://plato.stanford.edu/archives/win2025/entries/qualia-inverted/ (Accessed: 28 May 2026).

Cea, I., 2025. “Is Phenomenal Consciousness Necessary for AGI? A Review of the Theoretical Landscape”, *International Conference on Artificial General Intelligence*. Cham: Springer, pp. 60-71.

Chalmers, D.J., 1992. “Subsymbolic computation and the Chinese room”, in J. Dinsmore (Ed.), *The Symbolic and Connectionist Paradigms: Closing the Gap.* Hillsdale: Lawrence Erlbaum, pp. 25-48.

Chalmers, D.J., 1995. “Facing up to the Problem of Consciousness”, *Journal of Consciousness Studies*, 2(3), pp. 200-219.

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

Chalmers, D.J., 1997. “Moving forward on the problem of consciousness”, *Journal of Consciousness Studies*, 4(1), pp. 3-46.

Chalmers, D.J., 2003. “Consciousness and its place in nature”, in S. Stich & T. Warfield (Eds.), *The Blackwell Guide to Philosophy of Mind*. Oxford: Blackwell Publishing, pp. 102-142.

Chalmers, D.J., 2006. “Strong and Weak Emergence”, in P. Clayton & P. Davies (Eds.), *The Re-Emergence of Emergence: The Emergentist Hypothesis from Science to Religion.* Oxford: Oxford University Press, pp. 244-254.

Chalmers, D.J,. 2007. “The Hard Problem of Consciousness”, in S. Schneider & M. Velmans (Eds.) *The Blackwell Companion to Consciousness*. Oxford: Blackwell Publishing, pp. 225-235.

Chalmers, D.J., 2010. *The Character of Consciousness,* Oxford: Oxford University Press.

Chalmers, D.J., 2013. “How can we construct a science of consciousness?”, *Annals of the New York Academy of Sciences*, 1303(1), pp. 25-35.

Chalmers, D.J., 2015. “Panpsychism and panprotopsychism”, in T. Alter and Y. Nagasawa (Eds.), *Consciousness in the Physical World: Perspectives on Russellian Monism.* Oxford: Oxford University Press, pp. 102-154.

Chalmers, D.J., 2016. “The Combination Problem for Panpsychism”, in G. Bruntrup and L. Jaskolla (Eds), *Panpsychism: Contemporary Perspectives,* Philosophy of Mind Series. New York: Oxford Academic, pp. 179-214.

Chalmers, D.J., 2023. “Could a Large Language Model Be Conscious?”, *Boston Review*, 9 August. Available at: https://www.bostonreview.net/articles/could-a-large-language-model-be-conscious/ (Accessed: 01 June 2025).

Choi, S. & Fara, M., 2021. “Dispositions” in E.N Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2021 Edition). Stanford: Stanford University. Available at: https://plato.stanford.edu/archives/spr2021/entries/dispositions/ (Accessed: 27 September 2025).

Churchland, P.M., 1981. “Eliminative materialism and the propositional attitudes”, *The Journal of Philosophy*, 78(2), pp. 67-90.

Churchland, P. M., 1996. “The Rediscovery of Light”, *The Journal of Philosophy,* 93(5), pp. 211-228.

Churchland, P.M., & Churchland, P.S., 1981. “Functionalism, Qualia, and Intentionality”, *Philosophical Topics*, 12(1), pp. 121–145.

Churchland, P.M. & Churchland, P.S., 1990. “Could a machine think?”, *Scientific American*, 262(1), pp. 32-39.

Cleeremans, A. & Tallon-Baudry, C., 2022. “Consciousness matters: phenomenal experience has functional value”, *Neuroscience of consciousness,* 1, pp. 1-11.

Clifford, W.K., 1878. “On the nature of things-in-themselves”, *Mind*, 3(9), pp. 57-67.

Davidson, D., [1970]2001. “Mental events”, in *Essays on Actions and Events.* Oxford: Clarendon Press, pp. 207-228.

DeepSeek-AI, 2024. *DeepSeek-V3 Technical Report*. arXiv:2412.19437. Available at: https://arxiv.org/abs/2412.19437 (Accessed: 14 January 2026).

Dennett, D., 1979. “On the absence of phenomenology”, *Body, Mind, and Method: Essays in Honor of Virgil C. Aldrich.* Dordrecht: Springer, pp. 93-113.

Dennett, D., 1991. *Consciousness explained.* New York: Little, Brown and Company.

Dennett, D., 1992. *What Computers Still Can’t Do.* Cambridge: MIT Press.

Dennett, D., 2003. “Explaining the ‘magic’ of consciousness”, *Journal of Cultural and Evolutionary Psychology*, 1(1), pp. 7-19.

Dennett, D. C., 2004. “Can machines think?”, in S.M. Shieber (Ed.), *The Turing test: verbal behavior as the hallmark of intelligence*. Cambridge: MIT Press, pp. 269-292.

Descartes, R., [1641]1996. *Meditations on First Philosophy*. Edited by J. Cottingham. Cambridge: Cambridge University Press.

Dežman, D.V., 2024. “Promising the future, encoding the past: AI hype and public media imagery”, *AI and Ethics*, 4(3), pp. 743-756.

Dong, Y., Hou, J., Zhang, N., & Zhang, M., 2020. “Research on How Human Intelligence, Consciousness, and Cognitive Computing Affect the Development of Artificial Intelligence”, *Complexity*, 2020(1), pp. 1-11.

Dukor, M., 2014. “Philosophical Significance of Myths and Symbols: Zangbeto Cult”, *International Journal of Humanities, Social Sciences and Education*, 1(1), pp. 23–34.

Eddington, A.S., 1928. *The Nature of the Physical World*. Cambridge: Cambridge University Press.

Ewuoso, C. & Hall, S., 2019. “Core aspects of ubuntu: A systematic review”, *South African Journal of Bioethics and Law*, 12(2), pp. 93-103.

Flanagan, O. & Polger, T., 1995. “Zombies and the function of consciousness”, *Journal of Consciousness Studies,* 2, pp. 313-321.

Fodor, J.A., 1983. *The Modularity of Mind*. Cambridge: MIT Press.

Francescotti, R., 1999. “How to Define Intrinsic Properties”, *Noûs*, 33(4), pp. 590–609. Available at: http://www.jstor.org/stable/2671956 (Accessed: 20 September 2025).

Frankish, K., 2007. “The Anti-Zombie Argument”, *Philosophical Quarterly,* 57, pp. 650-666.

Frankish, K., 2022. “Illusionism is not trick”, *Human Affairs,* 32(3), pp. 321-327.

Gemini Team, Google, 2024. *Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context*. Google DeepMind. Available at: https://storage.googleapis.com/deepmind-media/gemini/gemini_v1_5_report.pdf (Accessed: 20 January 2026).

Gibb, S., 2015. “The Causal Closure Principle”, *The Philosophical Quarterly*, 65(261), pp. 626–647. Available at: https://doi.org/10.1093/pq/pqv030 (Accessed: 23 August 2024).

Goertzel, B., 2014. “Artificial general intelligence: concept, state of the art, and future prospects”, *Journal of Artificial General Intelligence*, 5(1), pp. 1-46.

Goff, P., 2009. “Why Panpsychism Doesn’t Help Us Explain Consciousness”, *Dialectica*, 63(3), pp. 289–311.

Goff, P., 2013. “Orthodox Property Dualism + Linguistic Theory of Vagueness = Panpsychism”, in R. Brown (Ed.), *Consciousness Inside and Out: Phenomenology. Neuroscience, and the Nature of Experience.* Dordrecht: Springer, pp. 75-91.

Goff, P., 2020. “Cosmopsychism, micropsychism and the grounding relation”, in W. Seager (Ed.), *The Routledge Handbook of Panpsychism*. London: Routledge, pp. 144–156.

Goff, P., 2024. “How exactly does panpsychism explain consciousness?”, *Journal of Consciousness Studies*, 31(3-4), pp. 56-82.

Goff, P., Seager, W., & Allen-Hermanson, S., 2022. “Panpsychism: The Philosophy of Mind”, in E.N. Zalta (Ed.), *Stanford Encyclopedia of Philosophy* (Summer 2022 Edition). Stanford: Stanford University. Available at: https://plato.stanford.edu/entries/panpsychism/ (Accessed: 16 September 2024).

Google DeepMind, 2025. *About Google DeepMind*. Available at: https://deepmind.google/about/ (Accessed: 13 January 2026).

Guyer, P. & Horstmann, R.P., 2023. “Idealism”, in E.N. Zalta & U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy* (Spring 2023 Edition). Stanford:

Stanford University. Available at: https://plato.stanford.edu/archives/spr2023/entries/idealism/ (Accessed: 30 January 2026).

Haikonen, P.O.A., 2013. “Consciousness and the Quest for Sentient Robots” in A. Chella, R. Pirrone, R. Sorbello & K. Jóhannsdóttir (Eds.), *Biologically Inspired Cognitive Architectures 2012. Advances in Intelligent Systems and Computing*, 196. Berlin: Springer.

Haugeland, J., 1985. *Artificial Intelligence: The Very Idea.* Cambridge: MIT Press.

Hendren, J., Grasso, M., Juel, B.E., & Tononi, G., 2024. “Overview: IIT as a Theory of Consciousness”, *IIT Wiki*. Center for Sleep and Consciousness, University of Wisconsin–Madison. Available at: https://www.iit.wiki/overview (Accessed: 15 October 2025).

Jackson, F., 1982. “Epiphenomenal Qualia”, *Philosophical Quarterly*, 32, pp. 127–136.

Jackson, F., 1986. “What Mary Didn’t Know”, *Journal of Philosophy*, 83(5), pp. 291–295.

Jackson, F., 2003. “Mind and Illusion”, *Royal Institute of Philosophy Supplement*, 53, pp. 251–271.

James, W., [1890]1918. *Principles of Psychology*. Cambridge: Harvard University Press.

Jin, I., Tangsrivimol, J.A., Darzi, E., Hassan Virk, H.U., Wang, Z., Egger, J., Hacking, S., Glicksberg, B.S., Strauss, M., & Krittanawong, C., 2025. “DeepSeek vs. ChatGPT: prospects and challenges”, *Frontiers in Artificial Intelligence*, 8, pp. 1-12.

Kim, J., 1990. “Supervenience as a philosophical concept”, *Metaphilosophy,* 21(1/2), pp. 1-27.

Kim, J., 1998. *Mind in a Physical World: An Essay on the Mind-Body Problem and Mental Causation*. Cambridge: MIT Press.

Kirk, R., 2023. “Zombies”, in E.N. Zalta & U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy* (Fall 2023 Edition). Stanford: Stanford University. Available at: https://plato.stanford.edu/archives/fall2023/entries/zombies/ (Accessed: 29 May 2026).

Koralus, P., 2024. *Why AI still isn’t getting into Oxford, and why that shows it is astounding*. Oxford HAI Lab. Available at:

https://hailab.ox.ac.uk/u/2024/10/Koralus2024WiredItalia.pdf (Accessed: 20 April 2025).

Kripke, S.A., 1980. *Naming and Necessity.* Cambridge: Harvard University Press.

Kuhn, R.L., 2024. “A landscape of consciousness: Toward a taxonomy of explanations and implications”, *Progress in Biophysics and Molecular Biology,* 190, pp. 28-169.

Kuhn, R.L., 2026. “What a ‘Landscape of Consciousness’ Means for Neurology and Neuroscience”, *Current Neurology and Neuroscience Reports,* 26(5), pp. 1-12.

Lange, F.A., 1925. *The History of Materialism and Criticism of Its Present Importance*. 1st edition. Translated by E.C. Thomas. London: Routledge.

Legg, S. & Hutter, M., 2007. “A Collection of Definitions of Intelligence”, in B. Goertzel & P. Wang (Eds.), *Advances in Artificial General Intelligence: Concepts, Architecture and Algorithms*. Amsterdam: IOS Press, pp. 17-24.

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

Levine, J., 1998. “Conceivability and the Metaphysics of Mind”, *Nous*, 32(4), pp. 449-480.

Lightman, H., Kosaraju, V., Burda, Y., Edwards, H., Baker, B., Lee, T., Leike, J., Schulman, J., Sutskever, I., & Cobbe, K., 2024. “Let's verify step by step”, *The Twelfth International Conference on Learning Representations (ICLR)*. Vienna, Austria, 7–11 May 2024. Available at: https://openreview.net/forum?id=v8L0pN6EOi (Accessed: 08 December 2025).

Maley, C.J., 2011. “Analog and digital, continuous and discrete”, *Philosophical Studies,* 115, pp. 117-131. Available at: https://doi.org/10.1007/s11098-010-9562-8 (Accessed: 17 June 2026).

Marquis, P., Papini, O. & Prade, H. (Eds.), 2020. *A Guided Tour of Artificial Intelligence Research: Volume I: Knowledge Representation, Reasoning and Learning*. Cham: Springer.

Mathews, F., 2009. “Why has the West failed to embrace panpsychism?”, in D. Skrbina (Ed.), *Mind that abides: Panpsychism in the new millennium*. Amsterdam: John Benjamins Publishing Company, pp. 341-360.

Mathews, F., 2020. “Living Cosmos Panpsychism”, in W. Seager (Ed.), *The Routledge Handbook of Panpsychism*. London: Routledge, pp. 131–143.

McCarthy, J., Minsky, M.L., Rochester, N., & Shannon, C.E., [1955]2006. “A proposal for the Dartmouth Summer Research Project on Artificial Intelligence, August 31, 1955”, *AI Magazine*, 27(4), pp. 12-14.

Millière, R. & Buckner, C., 2024*a*. *A Philosophical Introduction to Language Models – Part I: Continuity With Classical Debates*. arXiv preprint arXiv:2401.03910.

Millière, R. & Buckner, C., 2024*b*. *A Philosophical Introduction to Language Models – Part II: The Way Forward*. arXiv preprint arXiv:2405.03207.

Mitchell, M., 2019. “Artificial Intelligence Hits the Barrier of Meaning”, *Information*, 10(2), p. 51.

Mitchell, M., 2020. “On crashing the barrier of meaning in artificial intelligence”, *AI Magazine*, 41(2), pp. 86-92.

Mitchell, M., 2023. “Why AI is Harder Than We Think”, in C.F. Craver & C. Klein (Eds.), *Mind Design III: Philosophy, Psychology, and Artificial Intelligence*. Cambridge: MIT Press, pp. 175–188.

Mørch, H.H., 2017. “The evolutionary argument for phenomenal powers”, *Philosophical Perspectives*, 31, pp. 293-316.

Mørch, H.H., 2018. “Is the Integrated Information Theory of Consciousness Compatible with Russellian Panpsychism?”, *Erkenntnis*, 84(5), pp. 1065–1085.

Mørch, H.H., 2019. “Is consciousness intrinsic?: A problem for the integrated information theory”, *Journal of Consciousness Studies*, 26(1-2), pp. 133-162.

Mørch, H.H., 2020. “Does dispositionalism entail panpsychism?”, *Topoi*, 39(5), pp. 1073–1088.

Mørch, H.H., 2025. “Is Phenomenalism Superior to Russellian Panpsychism as a Theory of Consciousness and Matter?”, *Philosophia*, 52(2) pp. 487-497.

Müller, V.C., 2026. “Ethics of Artificial Intelligence and Robotics”, in E.N. Zalta & U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy* (Summer 2026 Edition). Stanford: Stanford University. Available at: https://plato.stanford.edu/archives/sum2026/entries/ethics-ai/ (Accessed: 20 June 2026).

Murray, S., 2016. “The Frame Problem”, in E.N. Zalta & U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy* (Spring 2016 Edition). Stanford: Stanford University. Available at: https://plato.stanford.edu/archives/spr2016/entries/frame-problem/ (Accessed: 19 June 2026).

Nagel, T., 1974. “What is it like to be a bat?”, *Philosophical Review*, 83, pp. 435–456.

Nagel, T., 1979, “Panpsychism”, in T. Nagel (Ed.), *Mortal Questions*, Cambridge: Cambridge University Press, pp. 181–195

Nagel, T., 1986. *The View From Nowhere*. Oxford: Oxford University Press.

Newell, A., & Simon, H. A., 1976. “Computer science as empirical inquiry: symbols and search”*, Communications of the ACM,* 19(3), pp. 113–126.

Nobel Prize Outreach, 2025. Press release. *NobelPrize.org*, 15 October. Available at: https://www.nobelprize.org/prizes/physics/2025/press-release/ (Accessed: 15 October 2025).

OpenAI, 2018. *OpenAI Charter*. Available at: https://openai.com/charter (Accessed: 14 January 2026).

OpenAI, 2025. *Reasoning guide*. Available at: https://platform.openai.com/docs/guides/reasoning (Accessed: 9 October 2025).

Papineau, D., 1993[2013]. *Philosophical Naturalism*. Available at: https://www.davidpapineau.co.uk/uploads/1/8/5/5/18551740/philosophical_naturalism_for_website_aug_2013.docx (Accessed: 30 May 2025).

Papineau, D., 2024. “Comments on Galen Strawson”, in G. Strawson & A. Freeman (Eds.), *Consciousness and its place in nature: why physicalism entails panpsychism*. 2nd edn. Exeter: Imprint Academic, pp. 129-140.

Penrose, R., 1994. *Shadows of the Mind*. Oxford: Oxford University Press.

Pinker, S., 1997. *How the mind works.* New York: W.W. Norton & Co.

Putnam, H., 1967. “Psychophysical Predicates”, in W. Capitan & D. Merrill (Eds.), *Art, Mind, and Religion.* Pittsburgh: University of Pittsburgh Press, pp. 158-167.

Putnam, H., 1992. “Why Functionalism Didn’t Work”, in J. Earman (Ed.), *Inference, Explanation, and Other Frustrations: Essays on the Philosophy of Science.* Berkeley: University of California Press, pp. 255-270.

Putnam, H., 1997. “Functionalism: Cognitive Science or Science Fiction?”, in D.M. Johnson & C. Erneling (Eds.) *The Future of the Cognitive Revolution.* Oxford University Press, New York, pp. 31–44.

Robinson, H., 2023. "Dualism", in E.N. Zalta & U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy* (Spring 2023 Edition). Stanford: Stanford University. Available at: https://plato.stanford.edu/archives/spr2023/entries/dualism/ (Accessed: 16 September 2025).

Robinson, W., 2023. “Epiphenomenalism”, in E.N. Zalta & U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy* (Summer 2023 Edition). Stanford: Stanford University. Available at: https://plato.stanford.edu/archives/sum2023/entries/epiphenomenalism/ (Accessed: 27 October 2025).

Russell, B., 1927. *The Analysis of Matter*. London: Kegan Paul, Trench, Trubner & Co.

Russell, B., 1959. *My Philosophical Development*. London: Routledge.

Russell, S.J., 2021. “Human-compatible artificial intelligence”, in S. Muggleton & N. Chater (Eds.), *Human-like machine intelligence*. Oxford: Oxford University Press, pp. 3-23. Available at: https://doi.org/10.1093/oso/9780198862536.003.0001 (Accessed: 19 April 2025).

Russell, S.J. & Norvig, P., 2022. *Artificial Intelligence: A Modern Approach*. 4th ed. Global Edition. Harlow: Pearson Education Limited.

Schade, M., 2017. “‘All a wi a wan’: Incarnation, the New Physics, Ubuntu”, *Journal of Arts, Science and Technology*, 10(1), pp. 140-155.

Schaffer, J., 2010. “Monism: The priority of the whole”, *Philosophical Review*, 119(1), pp. 31-76.

Seager, W., 1995. “Consciousness, information, and panpsychism”, *Journal of Consciousness Studies*, 2, pp. 272-288.

Seager, W., 2016. “Panpsychist Infusion”, in G. Brüntrup, & L. Jaskolla (Eds.), *Panpsychism: Contemporary Perspectives.* Oxford: Oxford University Press, pp. 229–248.

Seager, W., 2020. “Introduction: A Panpsychist Manifesto” in W. Seager (Ed.) *The Routledge Handbook of Panpsychism*. London: Routledge, pp. 1-12.

Searle, J.R., 1980. “Minds, Brains, and Programs”, *Behavioral and Brain Sciences*, 3(3), pp. 417–457.

Searle, J.R., 2002. “Why I am not a property dualist”, *Journal of Consciousness Studies,* 9(12), pp. 57-64.

Searle, J.R., 2008. “Reductionism and the Irreducibility of Consciousness”, in M.A. Bedau & P. Humphreys (Eds.), *Emergence: Contemporary Readings in Philosophy and Science.* Cambridge: MIT Press, pp. 69–80.

Seth, A.K., 2009. “Functions of consciousness” in W.P. Banks (Ed.), *Encyclopedia of Consciousness* (Vol 1). Oxford: Academic Press, pp. 279-293.

Seth, A.K., 2025. “Conscious artificial intelligence and biological naturalism”, *Behavioral and Brain Sciences,* pp. 1-42.

Seth, A.K. & Bayne, T., 2022. “Theories of consciousness”, *Nature Reviews Neuroscience*, 23(7), pp. 439-452.

Shapiro, L., 2021. “Elisabeth, Princess of Bohemia”, in E.N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2021 Edition). Stanford: Stanford University. Available at: https://plato.stanford.edu/archives/fall2021/entries/elisabeth-bohemia (Accessed: 19 May 2026).

Skrbina, D., 2005. *Panpsychism in the West*. Cambridge: MIT Press.

Skrbina, D., 2009. “Panpsychism in history: An overview”, in D. Skrbina (Ed.), *Mind that abides: Panpsychism in the new millennium*. Amsterdam: John Benjamins Publishing Company, 2009, pp. 1-29.

Skrbina, D., 2017. *Panpsychism in the West*. Cambridge: MIT Press.

Skrbina, D., 2020. “Panpsychism Reconsidered: A Historical and Philosophical Overview”, in W. Seager (Ed.) *The Routledge Handbook of Panpsychism*. London: Routledge, pp. 101-116.

Smart, J.J.C. 1959. “Sensations and Brain Processes”, *Philosophical Review,* 68, pp. 141-156.

Smart, J.J.C., 2017. “The Mind/Brain Identity Theory”, in E.N. Zalta and U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy* (Spring 2017 Edition). Stanford: Stanford University. Available at: https://plato.stanford.edu/archives/win2022/entries/mind-identity/ (Accessed: 20 June 2026).

Smuha, N.A. (Ed.), 2025. *The Cambridge Handbook of the Law, Ethics and Policy of Artificial Intelligence*. Cambridge: Cambridge University Press (Cambridge Law Handbooks).

Stoljar, D., 2024. “Physicalism”, in E.N. Zalta and U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy* (Spring 2024 Edition). Stanford: Stanford University. Available at: https://plato.stanford.edu/archives/spr2024/entries/physicalism/ (Accessed: 20 April 2025).

Strawson, G., 2006. “Panpsychism?: Reply to commentators with a celebration of Descartes”, *Journal of Consciousness Studies*, 13(10-11), pp. 184-280.

Strawson, G., 2008. “Realistic Monism: Why Physicalism Entails Panpsychism”, in *Real Materialism: and Other Essays*. Oxford: Oxford University Press, pp. 52–74. Available at: https://doi.org/10.1093/acprof:oso/9780199267422.003.0003 (Accessed: 21 October 2025).

Tononi, G., 2008. “Consciousness as integrated information: a provisional manifesto”, *The Biological Bulletin*, *215*(3), pp. 216-242.

Tononi, G., 2012. “Integrated information theory of consciousness: an updated account”, *Arch Ital Biol*, 150(2-3), pp. 56-90.

Tononi, G. & Koch, C., 2015. “Consciousness: here, there and everywhere?”, *Philosophical Transactions of the Royal Society B*, 370, pp. 1-18. Available at: http://dx.doi.org/10.1098/rstb.2014.016 (Accessed: 15 October 2025).

Turing, A.M., 1950. “Computing Machinery and Intelligence”, *Mind*, 59(236), pp. 433–460.

Tye, M., 2025. “Qualia”, in E.N. Zalta & U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy* (Fall 2025 Edition). Stanford: Stanford University. Available at: https://plato.stanford.edu/archives/fall2025/entries/qualia/ (Accessed: 25 May 2026).

Van Gulick, R., 2025. “Consciousness”, in E.N. Zalta & U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy* (Spring 2025 Edition). Stanford: Stanford University. Available at: https://plato.stanford.edu/archives/spr2025/entries/consciousness/ (Accessed: 13 October 2025).

Whitehead, A.N., [1927-28] 1960. *Process and Reality.* New York: Macmillan.