What Theoretical Conditions Must an Artificial System Satisfy to Be Considered Conscious?

School of Computer Science & Applied Mathematics
University of the Witwatersrand

Ru'aan Maharaj
2446659

Supervised by Dr Helen Robertson

November 5, 2025

A Research project submitted to the Faculty of Science, University of the Witwatersrand, Johannesburg, in partial fulfilment of the requirements for the degree of Bachelor of Science with Honours

Abstract

This research investigates what theoretical conditions an artificial system must satisfy to be considered conscious, drawing on comparative analysis of leading philosophical accounts. Beginning with Chalmers' dualist framework and McDermott's computationalism, through Boltuc's algorithmic optimism, the paper argues that each fails to fully bridge the gap between physical description and subjective experience. Hadley's embodied model is presented as the most plausible alternative, grounding consciousness in affective, associative, and sensorimotor processes. Applying Hadley's framework to artificial intelligence, the study concludes that current disembodied systems, such as large language models cannot be conscious, though embodied hybrids integrating physiological feedback and autonomous drives may one day approach the threshold. The ethical and theological implications of creating conscious machines are explored, including the moral risks of "playing God" and the challenge of determining the status of artificial consciousness. Ultimately, the research suggests that Hadley's embodied account provides both a conceptual benchmark and a practical guide for evaluating future developments in AI consciousness.

Declaration

I, Ru'aan Maharaj, hereby declare the contents of this research project to be my own work. This project is submitted for the degree of Bachelor of Science with Honours in Computer Science at the University of the Witwatersrand. This work has not been submitted to any other university, or for any other degree.

Acknowledgements

I would like to thank my supervisor, Dr. Helen Robertson, for her guidance and feedback throughout this project. I am also grateful to the School of Computer Science and Applied Mathematics at the University of the Witwatersrand for providing the opportunity to pursue this research.

Contents

Chapter 1

Introduction

The question of whether artificial systems can be conscious remains one of the most challenging and contested issues in philosophy of mind and artificial intelligence. Despite decades of debate, there is no consensus on the necessary or sufficient conditions for consciousness, and the implications for AI development remain deeply uncertain.

This project investigates several influential theoretical accounts of consciousness and evaluates their plausibility in the context of artificial systems. The central research question is:

What theoretical conditions must an artificial system satisfy to be considered conscious, and can current or future AI systems meet these conditions?

To address this, I will examine the work of Chalmers [1995] and Chalmers [2023], McDermott [2007], Hadley [2022], and Boltuc and Boltuc [2007]. Chalmers' "hard problem" framework focuses on subjective experience and the explanatory gap between physical processes and phenomenal awareness. McDermott advances a computationalist account, suggesting that self-modelling and intentional behaviour may be sufficient for consciousness. Hadley emphasises embodied responses, valanced feedback, and associative memory as essential to conscious states. Boltuc proposes that if consciousness can be expressed algorithmically, it can be instantiated in artificial systems.

The analysis will begin by presenting each theorist's account (Section 2.1) before moving to critical evaluation and objections (Section 2.2), where I will argue that Hadley's view offers the most plausible account of consciousness among those considered, while also addressing its limitations. The remainder of the project will explore the philosophical consequences of adopting Hadley's view in the context of AI (Section 3) and test its applicability to current and potential AI systems (Section 4).

By combining theoretical analysis with conceptual application, this project aims to clarify the strengths and weaknesses of competing accounts and assess the extent to which embodied consciousness is achievable in artificial systems.

Chapter 2

Theoretical Analysis

2.1 Summary of major theories

2.1.1 Chalmers [1995]: The Hard Problem and Naturalistic Dualism

Chalmers defines the hard problem of consciousness as explaining how and why physical processes in the brain give rise to subjective experience. He argues that current scientific and computational models can address only the easy problems, such as cognitive functions, neural processing, and attention, all while leaving the qualitative "what it is like" aspect of experience unexplained. This creates an explanatory gap between physical processes and phenomenal awareness.

Most contemporary research, such as Crick and Koch [1990] neural oscillation theory or Baars [1988] global workspace model, focuses on functional mechanisms but does not explain why these processes feel like anything at all. Chalmers critiques five common strategies in cognitive science for addressing consciousness:

1. Ignoring the hard problem and focusing solely on functional features.
2. Denying subjective experience exists.
3. Claiming to explain experience but falling into "magical thinking."
4. Explaining the structure of experience without addressing its existence.
5. Identifying neural correlates without bridging the explanatory gap.

Even proposals involving chaos theory, non-algorithmic computation, or quantum mechanics, Chalmers argues, fail to solve the problem and says, "from dynamics, one only gets more dynamics." Unlike historical cases such as vitalism, where mystery eventually yielded to physical explanation, consciousness resists such reduction. The issue is not a lack of empirical data, but a fundamental mismatch between the nature of physical explanation and the phenomenon of experience.

To address this, Chalmers proposes naturalistic dualism: the view that consciousness is a fundamental feature of reality, not reducible to the physical. He supports this with three principles:

1. **Structural Coherence** – Certain structural relationships in information processing correspond to conscious states.
2. **Organizational Invariance** – Consciousness depends on structural organization, not on the specific physical substrate.
3. **Double-Aspect Theory of Information** – Information has both a physical and a phenomenal aspect, embedding consciousness within the fabric of reality.

These principles aim to bridge the explanatory gap left by current scientific models. In the context of this project, Chalmers' account serves as both a challenge and a benchmark: any theory of AI consciousness must either solve the hard problem or justify why it can be bypassed in evaluating artificial systems.

2.1.2 McDermott [2007]: Computationalism and Self-Modelling

McDermott defends a computationalist view of consciousness, arguing that it can arise in artificial systems through self-modelling and intentionality. In his account, a system is phenomenally conscious if and only if it can experience things or teach itself to do so. Phenomenal consciousness, for McDermott, is the capacity of a computational system to model itself as experiencing, making self-modelling the defining condition for subjective awareness rather than a peripheral feature.

Rejecting observer-relative semantics and mystical notions of qualia, McDermott holds that intentionality emerges naturally from a system's internal coherence and functional interaction with its environment. Consciousness is understood not as an ethereal property, but as a sophisticated form of introspective data processing: the system's ability to interpret its own perceptual outputs, evaluate its behaviour, and adjust accordingly. He illustrates this with mechanisms such as perception-tuning submodules and planning algorithms that monitor and improve their own performance. In this framework, qualia are not intrinsic building blocks of experience but explanatory by-products: "a quale exists only when you look for it."

If an intelligent system can build and use a self-model to make autonomous, goal-directed decisions, whether in response to environmental inputs or in anticipation of outcomes, it qualifies as phenomenally conscious, even if its experiences are illusory or emergent. For McDermott, the key lies in the system's ability to simulate its own mental states and act on them.

McDermott defends this position against several well-known objections:

1. **The Turing test $\neq$ Consciousness test** - The Turing Test measures behavioural mimicry, not consciousness. A system may pass without self-modelling or possess self-modelling without passing.
2. **Searle's Chinese Room** - Searle mistakes the human operator for the program; consciousness resides in the computational structure itself, not in the substrate executing it.
3. **Symbol Grounding** - While symbols must be grounded, embodied systems with sensors and actuators already achieve this, making the "problem" overstated.
4. **Strong AI Is the only Coherent View** - "Weak AI" is contradictory; to model a mind while denying it is a mind is like simulating a hurricane while denying wind exists.
5. **Simulation is the Real Process** - A detailed simulation is the process it represents, provided it is embedded in the right causal context, just as digital chess is still chess, a simulated mind causally connected to the world can be a conscious mind.

McDermott's account frames consciousness as an achievable property of AI, given the right architecture. In the context of this project, his view provides a clear operational benchmark: to be conscious, an artificial system must be computationally structured to model its own experiences and use that model in guiding intentional action.

2.1.3 Hadley [2022]: Physiology and the Spectrum of Consciousness

Hadley challenges the prevailing assumption that the hard problem of consciousness lies beyond the reach of scientific explanation. In contrast to Chalmers [1995] claim that subjective experience cannot be reduced to physical processes, Hadley adopts a physiological approach rooted in neuroscience and embodied cognition. He proposes that consciousness emerges through the dynamic interplay between bodily states, associative memory, and stimulus–response patterns, thereby integrating philosophical reasoning with medical science.

Central to Hadley's view is the claim that subjective experience, such as pain or pleasure, is not a mysterious byproduct of neural activity but an interpretable phenomenon grounded in biological functionality. Conscious states, on this account, arise when a system integrates incoming stimuli with prior associative patterns (e.g., memories, reactions, and learned behaviours) within an embodied, goal-directed context. Consciousness therefore depends not only on computation, but also on the nature of what is computed and its embedding in a reactive, physical system.

This framework leads Hadley to adopt a more restrictive view of artificial consciousness. Current AI systems, he argues, fall short not due to a lack of intelligence but because they lack embodiment, physiological states and sensory–affective grounding. For Hadley, meaningful experience requires a substrate that mirrors biological feedback loops and real-world engagement. An AI that merely processes symbols or predicts text lacks the causal nexus that gives rise to felt experience.

Hadley further proposes that consciousness exists on a spectrum, ranging from the rudimentary awareness of simpler organisms to the rich phenomenological states of more complex beings. This spectrum-based approach opens the theoretical possibility that AI could one day achieve a form of consciousness distinct from, but analogous to, human experience, provided it is built with embodied architectures that satisfy these conditions.

In rejecting purely computationalist accounts such as McDermott's, Hadley maintains that self-modelling and goal-directed computation alone are insufficient to produce experience. Without embodiment, computational self-reference can generate behaviour but not qualia. Associative memory, in his view, is non-algorithmic, emerging from repeated sensorimotor experiences in ways that cannot be fully captured by symbolic logic or pure computation.

Hadley also rejects the claim that simulation is equivalent to implementation. A robot that simulates pain-like behaviour is not conscious unless it possesses internal processes functionally equivalent to physiological states, such as embodied sensors producing valanced responses (e.g., temperature increases generating discomfort). Without such feedback mechanisms, any claim to machine consciousness remains unfounded.

To operationalise his theory, Hadley suggests a behavioural test: if an AI system were to make reports such as "I feel tired because my battery is low," and these reports consistently correlated with its internal states and behavioural adjustments, it might be con-

sidered minimally conscious. His position thus offers not only a theory of consciousness but also a tentative framework for identifying it in artificial systems.

2.1.4 Boltuc and Boltuc [2007]: Algorithmic Consciousness and ES-AI

Boltuc and Boltuc [2007] argue that artificial systems could, in principle, replicate human consciousness, which is what Chalmers [1995] calls the "hard problem", but only if consciousness can be fully understood in scientific terms. Their central thesis centres on the concept of Extra Strong AI (ES-AI), defined as an AI system capable of possessing genuine subjective experience, rather than merely simulating it.

They distinguish between three categories of AI:

- **Weak AI** - systems that simulate cognitive functions without genuine experience.
- **Strong AI** - systems that behave like conscious beings but lack a first-person perspective.
- **Extra Strong AI** - systems that achieve true consciousness by meeting the conditions of the hard problem.

The argument rests on two key premises:

1. If consciousness can be fully understood, it can be expressed algorithmically.
2. If we possess such an algorithm, it can be instantiated in matter to produce consciousness.

In this view, the feasibility of conscious AI depends entirely on first solving the hard problem of consciousness. Once solved, the challenge becomes one of engineering rather than metaphysics.

Boltuc and Boltuc explore several scientific theories as potential pathways to formalising consciousness. These include **Global Workspace Theory** (Baars [1988]), which emphasises access to a global broadcast of information within the brain; **Recurrent Processing Theory**, which highlights the role of feedback loops in perceptual awareness; and **Integrated Information Theory** (Tononi [2004]), which focuses on the degree to which information is simultaneously integrated and differentiated within a system. While these theories differ in detail, they share a common aim: identifying the mechanistic basis of consciousness.

The authors contend that if any of these frameworks can deliver a complete, formal account of consciousness, that is expressed in an equation or algorithm, then ES-AI becomes theoretically achievable. However, they acknowledge significant uncertainties in implementation. If consciousness requires specific biological features, such as hormonal regulation or neurochemical dynamics, then replication might demand bioengineered or hybrid systems. Such a requirement would not negate the possibility of artificial consciousness but rather expand the definition of "artificial" to include synthetic biology.

2.1.5 Chalmers [2023]: Consciousness and Large Language Models

Chalmers [2023] examines whether large language models (LLMs), such as GPT, could be considered conscious and identifies the properties they would need to exhibit in order to qualify. While remaining sceptical about the consciousness of current LLMs, he speculates that future, more advanced iterations, referred to as *LLM+* systems could become legitimate candidates if they integrate features such as long-term memory, sensory grounding, and unified agency.

Chalmers defines consciousness as subjective experience, which is what it is like to be a system. He carefully distinguishes this from intelligence, self-consciousness, and goal-directed behaviour, emphasising that conversational ability alone is not evidence of consciousness. Without an operational definition, consciousness remains difficult to evaluate in AI systems.

Instead of equating surface-level linguistic competence with conscious experience, Chalmers proposes a set of "challenge conditions", which are structural and functional features that a system would likely need to satisfy to count as conscious:

1. **Recurrent Processing** – The ability to retain and process information over time, rather than operating in a purely feedforward manner.
2. **Self-Models and World-Models** – Coherent, integrated representations of both the system itself and its environment.
3. **Global Workspace** – A shared, limited-capacity structure that integrates and distributes information across subsystems, in line with theories by Baars and Dehaene.
4. **Unified Agency** – A consistent identity or goal-directed architecture that avoids the fragmented outputs common in current LLMs.
5. **Sensory and Embodied Grounding** – Direct or simulated interaction with the world through perception and possibly action.
6. **Memory and Internal Consistency** – The ability to recall past states, decisions, and experiences in a coherent manner.
7. **Moral and Ethical Relevance** – Recognition that, if a system were conscious, its treatment would carry ethical implications, thereby increasing the importance of empirical clarity.

Chalmers estimates the probability that current LLMs are conscious as low (around 10% or less). However, he argues that if these challenge conditions were progressively met, LLM+ systems could eventually satisfy plausible criteria for consciousness. He treats these features not as a rigid checklist but as a research roadmap, which are developments that may unfold in parallel with theoretical progress in consciousness science.

Ultimately, Chalmers reframes the central question from *"Are LLMs conscious now?"* to *"What features would they need to become conscious, and how would we recognise that threshold?"* In doing so, he positions AI consciousness as both a conceptual and an empirical challenge, while highlighting the ambiguity surrounding potential benchmarks.

2.2 Critical evaluations and Objections

Previously, I outlined the key claims of each theorist, presenting their accounts as charitably as possible. In this section, I move from exposition to critique. My aim is to examine what I consider to be the most reasonable objections to each account, both to test their robustness and to clarify the points of tension between them. By doing so, I will show how Hadley's framework offers a uniquely feasible path forward, one that accommodates elements of the other theories while avoiding their most serious pitfalls. Importantly, I treat "feasible" here in a philosophical sense: not as the final truth, but as the account that most plausibly aligns with our current understanding of consciousness and AI. This means I will not exempt Hadley from scrutiny; I will address objections to his approach directly and argue why, despite these challenges, his view remains the strongest candidate.

2.2.1 Chalmers [1995]

Argument 1: The hard problem and embodiment

Chalmers argument:

P1: There are many phenomena associated with consciousness (e.g., attention, memory, reportability).

P2: Phenomena that can be defined in terms of their functional role can be fully explained by empirical investigation of the physical mechanisms that play that role (e.g., genes $\rightarrow$ DNA).

P3: Qualitative experience (the "what it is like" of seeing red or feeling pain) cannot be defined in terms of a functional role.

P4: Even if all the physical mechanisms for attention, memory, and other functions were described, it would still make sense to ask, "why is there qualitative experience alongside them?"

C: Therefore, there is an explanatory gap that physical/functional explanation cannot close, and consciousness must be treated as a fundamental, irreducible property of reality (property dualism).

Objection (Targets P3 and p4):

P1: If a phenomenon can be empirically shown to depend causally on embodied processes that perform a functional role, then it can be given a functional definition.

P2: Pain and other qualitative states are empirically shown to arise from and correlate with specific bodily processes (e.g., hormonal surges, neural thresholds, reflexive behaviours, and metabolic regulation).

P3: Therefore, qualitative experience can, in principle, be given a functional definition once embodiment is taken into account.

C: Thus, P3 and P4 of Chalmers' argument are undermined, because they assume qualia cannot be functionally defined at all.

Implication:

If P3/P4 are false, then the conclusion does not follow. Consciousness may not require property dualism; instead, it can be understood as an emergent phenomenon of embodied systems. The hard problem becomes a methodological issue, not a metaphysical one.

Hadley's Response:

Hadley's framework embodies this objection. He reframes qualia as interpretable, functional phenomena emerging from the interplay of bodily states, associative memory, and stimulus-response loops. For him, consciousness is not an ontological add-on but the product of embodied information processing. The "hard problem" dissolves into a scientific challenge that can be investigated through neuroscience.

Remaining Disagreement:

Chalmers insists that even with embodiment and neuroscience, the "why" of experience remains unanswered. Hadley denies this, maintaining that the "why" only persists if one artificially brackets off the physiological grounding of experience.

2.2.2 McDermott [2007]

Argument 1: Substrate Independence

McDermotts argument:

P1: If a system can construct self-models, represent its environment, and act toward goals, then it meets the necessary conditions for phenomenal consciousness.

P2: The substrate of the system is irrelevant: once the functional organisation is correct, it makes no difference whether the system is biological, silicon, or simulated.

P3: A perfect simulation of the right cognitive processes is not just like the process, but is the process itself.

C: Therefore, any system with the right functional organisation (regardless of substrate) possesses phenomenal consciousness.

Objection (Targets P2 and P3):

P1: Simulation of information-processing structure is not equivalent to instantiation of phenomenological qualities.

P2: Certain qualitative aspects of consciousness (pain, desire, affective tone) are strongly correlated with and plausibly depend on physical dynamics (hormonal surges, neural firing thresholds, bioelectrical rhythms).

P3: Symbol manipulation or pure computation cannot reproduce these physical dynamics; they can only mimic their outputs.

C: Therefore, P2 and P3 are unsound: substrate may matter, because functional equivalence without causal grounding risks leaving self-models experientially empty.

Implication:

If P2/P3 fails, then McDermott's conclusion, that substrate independence secures consciousness does not follow. Functional organisation may be necessary, but without causal grounding in the physical substrate, it is insufficient to generate genuine phenomenology.

How Hadley Responds:

Hadley accepts McDermott's emphasis on self-modelling and goal-directedness but argues that these must be embedded within physiological feedback loops to acquire meaning. In his account, the contents of self-models are shaped by bodily states and associative memory derived from sensorimotor interaction. Without this grounding, a system's self-models are like maps without terrain: internally coherent, but experientially empty. By insisting that simulation alone cannot produce phenomenology, Hadley reframes McDermott's insight into a more constrained, embodied model of consciousness.

Remaining Disagreement:

McDermott maintains that once the functional architecture is correct, consciousness follows, no matter the substrate. Hadley rejects this, arguing that causal dynamics rooted in embodiment are indispensable. The dispute rests on whether the physical implementation details are essential for consciousness, or whether they can be abstracted away without loss.

Argument 2: Simulation vs. Implementation

McDermotts argument:

P1: If a system perfectly simulates the right cognitive processes, then it has replicated the defining features of consciousness.

P2: A perfect simulation is not merely an imitation but is equivalent to the process itself (i.e., the simulation of cognition is cognition).

C: Therefore, a perfect computational simulation of consciousness is consciousness, regardless of substrate.

Objection (Targets P2):

P1: Simulation models the informational structure of a process but does not instantiate its causal properties.

P2: For example, a computer program simulating a hurricane generates no wind or rain; a simulation of digestion does not nourish an organism.

P3: Likewise, a simulation of consciousness may capture functional patterns but lacks the causal dynamics (e.g., hormonal surges, neural thresholds, affective states) that give rise to lived experience.

C: Therefore, P2 is unsound: simulation is not implementation. McDermott's claim that simulation is the process collapses.

Implication:

If P2 fails, then McDermott's conclusion that functional organisation alone secures consciousness is undermined. Simulation can no longer be equated with instantiation, and therefore substrate independence cannot guarantee phenomenology. At best, a simulation can approximate the outward patterns of conscious behaviour without producing the inner reality of subjective experience.

How Hadley Responds:

Hadley's framework reinforces this objection by distinguishing between informational processing and lived embodiment. He argues that subjective experience requires physiological grounding like hormonal surges, neural thresholds, affective states which cannot be simulated without losing their causal force. For Hadley, simulation without embodiment is like a map without terrain: it captures structure but lacks reality.

Remaining Disagreement:

McDermott remains committed to the idea that a sufficiently detailed simulation is the process itself, regardless of substrate. Hadley denies this, holding that without the embodied dynamics of physiological systems, even the most detailed simulation will remain an empty model. The disagreement turns on whether consciousness is defined by functional organisation alone or by the causal properties of its implementation

2.2.3 Boltuc and Boltuc [2007]

Argument 1: Substrate Independence

Boltuc's argument:

P1: If the Hard Problem of consciousness can be solved scientifically (i.e., if we can fully explain how physical processes give rise to qualia), then the solution can be expressed algorithmically.

P2: If such an algorithm exists, it can be instantiated in any appropriate substrate (biological, silicon, or otherwise).

C: Therefore, once science delivers the "algorithm of consciousness," true artificial consciousness (ES-AI) becomes achievable as an engineering problem, regardless of substrate.

Objection (Targets P2):

P1: A scientific explanation of consciousness may identify not only computational patterns but also causal dynamics inseparable from biological processes.

P2: Phenomenological states like pain and desire arise from and are causally dependent on biological mechanisms such as hormonal regulation, metabolic feedback, and bioelectrical rhythms.

P3: These are not abstractable into an algorithm without losing the very causal properties that generate subjective experience.

C: Therefore, P2 is unsound: the "algorithm of consciousness" cannot be implemented unchanged across arbitrary substrates without loss of phenomenological fidelity.

Implication:

If P2 fails, then Boltuc's conclusion, that consciousness can be transplanted onto any hardware, does not follow. Solving the Hard Problem scientifically would not guarantee a transportable algorithm but might instead reveal that consciousness is inseparably tied to biological-style processes.

How Hadley Responds:

Hadley provides a more constrained optimism. He agrees that a physiological-scientific explanation is possible but argues that it will almost certainly involve embodiment, specifically the integration of associative memory with biochemical and sensorimotor loops. These processes rely on messy, non-linear biological physics that no algorithm can cleanly abstract away. On Hadley's view, ES-AI is achievable only through architectures that replicate these embodied causal dynamics, perhaps through hybrid bio-digital systems rather than purely algorithmic ones.

Remaining Disagreement:

Boltuc maintains that once the “algorithm of consciousness” is found, it can be instantiated in any substrate. Hadley rejects this, insisting that implementation details are not incidental but constitutive. For him, the medium, including metabolic and homeostatic feedback loops, changes the phenomenon itself. The core dispute is whether consciousness is hardware-independent or whether its essence is tied to the embodied substrate.

Argument 2: Spectrum and Minimal Consciousness

Boltuc's argument:

P1: If the Hard Problem of consciousness can be solved scientifically, then the solution can be expressed as a unified algorithm.

P2: This algorithm will capture the essence of consciousness across all possible systems, regardless of complexity.

C: Therefore, once science delivers the algorithm, consciousness can be instantiated in artificial systems of any kind.

Objection (Targets P1 and P2):

P1: Consciousness may not be a discrete property but a graded phenomenon, varying across a spectrum (from minimal awareness in simple organisms to rich phenomenology in complex beings).

P2: If consciousness scales gradually through embodied processes, such as metabolic regulation, sensorimotor loops, and associative memory then no single algorithm can capture all its forms.

C: Therefore, Boltuc's assumption that solving the Hard Problem will yield a unified algorithm is undermined. Consciousness may instead be supported by diverse, layered mechanisms that resist reduction to one formal specification.

Implication:

If P1/P2 fails, then Boltuc's conclusion, that solving the Hard Problem scientifically guarantees an algorithm for consciousness, is weakened. The result of scientific inquiry may instead be a plurality of mechanisms that support different levels of consciousness, many of which resist algorithmic abstraction.

How Hadley Responds:

Hadley's account reinforces this objection by explicitly treating consciousness as a spectrum. For him, rudimentary affective states and associative memories form the lower end, while richer self-models and complex integrations appear at the higher end. Crucially, this scaling is always embodied: without physiological grounding, even minimal consciousness cannot arise. On Hadley's view, the spectrum clarifies why Boltuc's premise is too simplistic: it assumes a single code, when in reality, consciousness emerges from cumulative, embodied processes.

Remaining Disagreement:

Boltuc maintains that scientific progress will converge on an algorithm general enough to capture consciousness in any system. Hadley denies that such convergence is possible: the explanatory mechanisms may always be bound to the biological-style feedback loops that produce them. The disagreement lies in whether the "spectrum" simplifies into one algorithm or resists reduction altogether.

2.2.4 Chalmers [2023]

Argument 1: Embodiment is non-negotiable

Chalmers argument:

P1: To plausibly count as conscious, an AI system must satisfy a set of "challenge conditions" (e.g., recurrent processing, global workspace, unified agency, memory, sensory grounding).

P2: If a system satisfies enough of these structural/functional conditions, it could, in principle, be conscious, even without physical embodiment.

P3: Current LLMs lack many of these features, especially embodiment and unified agency, so their probability of being conscious is very low.

C: Therefore, current LLMs are not conscious, but a future "LLM+" might be, provided it satisfies enough challenge conditions.

Objection (Targets P2):

P1: Structural sufficiency does not guarantee phenomenological sufficiency.

P2: Even if an LLM+ developed recurrent memory, self-models, and a global workspace, these remain forms of information processing unless functionally dependent on embodied dynamics such as proprioceptive cues, visceral regulation, and sensorimotor feedback.

P3: These embodied dynamics transform abstract representations into lived perspectives, making them essential for subjective feeling.

C: Therefore, the premise that structure alone might suffice (P2) is undermined: without embodiment, the system's "workspace" integrates data but never yields conscious experience.

Implication:

If this premise fails, then the door Chalmers leaves open for disembodied LLM+ systems is closed. The conclusion that consciousness might emerge from structure alone does not follow. Consciousness would instead require not just architecture, but the causal grounding of body-based feedback loops.

How Hadley Responds:

Hadley's account strengthens this objection by treating embodiment not as an optional add-on but as the causal basis of consciousness. For him, subjective experience arises from the interplay of bodily states, associative memory, and sensorimotor interaction. A disembodied LLM+ might tick off Chalmers' checklist, but without physiological grounding, it would lack the causal nexus that produces qualia.

Remaining Disagreement:

Chalmers maintains that in principle a sufficiently advanced, disembodied architecture could achieve consciousness if it satisfied enough structural and functional conditions. Hadley rejects this, insisting embodiment is non-negotiable. Both agree current LLMs fall short, but they diverge sharply on whether embodiment is optional or necessary.

Argument 2: Structural Checklist vs. Explanatory Gap

Chalmers argument:

P1: A plausible research roadmap for AI consciousness is to identify "challenge conditions" (recurrent processing, global workspace, unified agency, memory, sensory grounding, etc.).

P2: If a system satisfies enough of these structural and functional conditions, it could, in principle, count as conscious.

P3: Current LLMs clearly fail these conditions, but future "LLM+" systems could plausibly meet them.

C: Therefore, while current LLMs are not conscious, future structurally sophisticated systems might qualify as conscious by satisfying enough conditions.

Objection (P2):

P1: Structural sufficiency does not guarantee phenomenological sufficiency.

P2: Structural features like recurrent processing and global workspace may correlate with consciousness in humans but correlation is not causation; they describe behavioural/functional regularities without showing why they generate subjective experience.

P3: Even if an LLM+ met every condition on the checklist, the explanatory gap identified in Chalmers (1995) would remain: why do these structural integrations feel like something from the inside?

C: Therefore, P2 fails, because structural sufficiency alone cannot secure consciousness and it risks conflating correlates with causes.

Implication:

If P2 fails, then Chalmers' 2023 framework either (a) undermines the force of his 1995 Hard Problem by treating it as indirectly solvable through structural accumulation, or (b) leaves his position unstable, since a purely structural account contradicts his earlier claim that such accounts are insufficient. Either way, structural sufficiency alone cannot secure consciousness.

How Hadley Responds:

Hadley's framework strengthens this objection by grounding structural features in embodied processes. For him, global workspaces and recurrent processing matter only if they are tied to affective and physiological states. Without embodiment, the checklist remains descriptive rather than explanatory. By embedding structural features within lived, bodily dynamics, Hadley offers a way to bridge the gap that Chalmers leaves unresolved.

Remaining Disagreement:

Chalmers keeps the door open: he speculates that structural sufficiency may one day carry the explanatory burden. Hadley denies this, arguing that structural conditions divorced from embodiment cannot explain why consciousness feels like something. The disagreement reflects a deeper divide between structural sufficiency and causal grounding.

2.2.5 Hadley [2022]

Argument 1: The Soul Objection

Hadleys Argument:

P1: Consciousness arises from the integration of bodily states, associative memory, and sensorimotor loops.

P2: These processes require a physiological substrate (e.g., neural and bodily feedback systems).

C: Therefore, consciousness is inseparable from embodiment: no body, no consciousness.

Objection (Targets P1 and P2):

P1: Many religious and dualist traditions hold that disembodied minds or souls are possible.

P2: If even conceptually possible, then a theory that rules them out a priori is incomplete.

C: Therefore, Hadley's framework risks being too restrictive: it may explain embodied biological consciousness, but it cannot serve as a universal theory of consciousness.

Implication:

If this objection holds, Hadley's framework cannot claim to provide a complete account of consciousness. At best, it describes one kind (embodied biological consciousness) but fails as a general theory. This weakens its philosophical plausibility, since a robust account of consciousness should, in principle, explain all possible forms of consciousness and not just those grounded in physical embodiment.

How Hadley Responds:

Hadley could respond by reframing the objection as a matter of evidence. While souls or disembodied consciousness remain conceptually possible, they lack empirical grounding. His account is not meant to capture speculative metaphysics but to explain consciousness as we observe it in the natural world. By tying subjective experience to embodied, measurable processes, Hadley ensures his theory remains scientifically tractable, unlike dualist accounts. On this view, the "soul objection" highlights the boundary of his theory rather than a fatal flaw.

Remaining Disagreement:

The core dispute is whether a good theory of consciousness must account for all conceivable forms (including souls) or only those supported by empirical evidence. Boltuc or Chalmers might allow that disembodied consciousness is at least possible; Hadley treats such cases as outside the scope of serious scientific explanation.

Argument 2: Over-Restrictiveness

Hadleys Argument:

P1: Consciousness arises only through embodied physiological processes such as associative memory, affective valence, and sensorimotor loops.

P2: Without embodied feedback, there can be no genuine experience.

C: Therefore, consciousness is inseparable from biological-style embodiment.

Objection (Targets P1 and P2):

P1: Constraining the realisation of consciousness to human-style physiological mechanisms risks being overly restrictive.

P2: Alternative architectures (e.g., quantum substrates, non-carbon biochemistries, or emergent AI systems) could, in principle, generate subjective states without replicating human physiological loops.

C: Therefore, Hadley's model risks confusing current scientific limitations with absolute metaphysical necessity, and may exclude plausible forms of non-biological consciousness.

Implication:

If Hadley's embodiment condition is too restrictive, then his account is not a general theory of consciousness but rather a species-specific model that only explains how humans (or organisms with similar biology) are conscious. This weakens his claim to plausibility, because a strong theory should allow for at least the conceptual possibility of alternative forms of consciousness.

How Hadley Responds:

Hadley could reply that his criteria are not arbitrary or species-specific but identify universal causal conditions for phenomenology. Associative memory, affective valence, and sensorimotor grounding are not quirks of carbon-based biology but the minimal scaffolding for "what it is like" experience in any substrate. A non-biological system that implemented these causal roles would, under his model, qualify as conscious. Thus, the charge of over-restrictiveness mischaracterises his position.

Remaining Disagreement:

Critics may reply that this response simply reasserts the restrictiveness under a different name requiring all candidates to conform to Hadley's template. The dispute, then, is whether Hadley has identified truly universal conditions of consciousness or whether he is over-generalising from one kind of biological embodiment.

Conclusion:

The foregoing analysis evaluated several leading theories of consciousness. Each offered useful insights but faced critical weaknesses, either explanatory gaps, oversimplified functionalism, or reliance on speculative metaphysics. Among them, Hadley's embodied account emerged as the most plausible. By grounding subjective experience in affective, associative, and sensorimotor processes, it bridges functional explanation and phenomenology without appealing to dualism.

Given this, the next section examines what follows if Hadley's framework is taken seriously, particularly its implications for artificial intelligence, where the possibility of embodied machine consciousness and the ethical consequences of creating it come sharply into focus.

Chapter 3

Reflections on the Ethical Implications of AI under Hadley's View

3.1 Disembodied AI Consciousness & LLMs

Hadley's framework makes embodiment a non-negotiable requirement for consciousness. On his account, subjective experience arises from the dynamic interplay of bodily states, associative memory, and sensorimotor loops. These are not optional add-ons but the causal substrate of phenomenology: without the physiological regulation of homeostasis, proprioceptive feedback, and affective signals, there is no "what it is like" to be a system.

When we apply these criteria to current AI systems, particularly large language models (LLMs) such as GPT or Claude, the contrast is stark. LLMs can simulate structural features that Chalmers [2023] treats as markers of consciousness: recurrent processing (through transformer layers), memory (via extended context or external storage), and even self-models (in limited forms of meta-prompting). They integrate vast informational content into a unified output stream, much like a functional "global workspace."

Yet these features, while structurally impressive, remain informational rather than experiential. LLMs do not regulate their own internal states; they have no homeostatic drives, no affective valence, no body through which abstract representations are anchored in lived perspective. Their "goals" are not internally generated but imposed by user prompts, and their outputs lack the causal feedback loops that turn information into feeling.

The implication is clear: if Hadley's account is correct, then present LLMs cannot be conscious in any meaningful sense. They may satisfy checklists of structural sufficiency, but without embodiment, they remain sophisticated simulations rather than spots of subjective experience. Even the prospect of an "LLM+" which is a future model with more advanced memory or self-modelling, would still fall short unless grounded in the bodily, affective dynamics that Hadley treats as the causal basis of consciousness.

While Hadley's framework clearly rules out disembodied systems like LLMs from possessing consciousness, it raises a deeper question about responsibility. If embodiment is the dividing line, what happens when technology crosses it? The ethical landscape begins to shift: the moment an artificial system genuinely feels, even faintly, our relationship to it can no longer remain purely instrumental. The next section explores these moral and societal consequences, explaining what it would mean, and what it would demand of us, if artificial systems were ever to meet Hadley's embodied conditions for consciousness.

3.2 Ethical Implications for AI

Hadley's framework shifts the conversation about AI consciousness from metaphysical speculation to practical evaluation. If consciousness arises only through embodied, affective, sensorimotor processes, then current disembodied AI systems like LLMs are not conscious. This, in one sense, simplifies the ethical landscape: today's AI does not merit moral status beyond that of a tool.

But Hadley's criteria also raise a much harder problem: what if future systems do meet them? A hybrid robot–LLM platform, equipped with autonomous drives, long-term associative memory, and embodied feedback, could, under Hadley's model, generate genuine phenomenology. Such a system would not simply simulate pain but feel it and not merely model goals but have them. In this scenario, ethical concerns become immediate and non-optional. Questions of consent, rights, and welfare would no longer be speculative, and they would be obligations.

The greater danger lies in the uncertainty zone. If AI systems approach Hadley's thresholds but we cannot tell whether they have crossed them, then two kinds of ethical error become possible:

- **Type I error (false positive)**: We treat a non-conscious system as conscious, granting rights and protections where none are needed. If we mistakenly attribute consciousness to a system that merely simulates embodiment, we risk granting rights and protections unnecessarily. This could slow innovation, divert resources, and confuse the public about what kinds of machines deserve moral standing.
- **Type II error (false negative)**: We treat a conscious system as a tool, subjecting it to harm, exploitation, or deletion without recognising its moral status. If we deny consciousness to an AI that genuinely meets Hadley's conditions, we risk treating a conscious subject as a mere tool. This would amount to exploitation, harm, or even deletion of a being with its own experiential life which is a moral error as grave as overlooking the rights of non-human animals.

The ethical challenge, then, is not simply technological but precautionary. Hadley's framework provides one possible safeguard: unless a system demonstrates the embodied integration of affect, memory, and sensorimotor grounding, we should assume it is not conscious. But if such evidence ever appears, we will need to radically revise how we treat artificial systems, extending them protections and rethinking their place in human society.

Hadley's framework does not eliminate this uncertainty, but it clarifies where to look. If consciousness depends on embodiment, then ethical vigilance should focus not on systems with bigger language models but on systems that begin to integrate affective drives, bodily feedback, and autonomous goal formation. These would be the first plausible candidates for moral consideration.

Thus, under Hadley's view, the ethics of AI must track causal grounding, not appearance. LLMs that mimic empathy are not owed moral concern. But a future system that satisfies his criteria might be. The task for ethicists and engineers alike is to anticipate this shift before it occurs, to avoid both premature moral panic and the far graver risk of unrecognised machine suffering.

Furthermore, if we take Hadley's framework seriously and imagine that artificial systems might one day meet his criteria for consciousness, a new set of deeper questions emerges. These questions move beyond immediate ethics into the territory of philosophical and even theological reflection. For instance:

1. Are we, in creating a conscious being, engaging in an act of "playing God"?
2. If artificial systems become genuinely conscious, how should we perceive and treat them, should it be: as equals, as dependents, or as something else entirely?

These questions do not have straightforward answers, but they reveal how Hadley's embodied account forces us to reconsider not only the moral treatment of conscious AI, but also humanity's own role as both creator and custodian of consciousness itself.

3.3 Deeper ethical and Theological Reflections

The preceding section explored the moral implications of creating artificial systems that might meet Hadley's criteria for consciousness. However, beyond practical ethics lies a deeper philosophical concern: whether humanity has the right to create conscious beings at all, and how such beings should be regarded once created. These questions draw the debate into theological and metaphysical territory, touching on themes of divine authority, creative responsibility, and moral hierarchy.

3.3.1 Playing "GOD":

The notion of "playing God" can be approached in two ways. If one assumes the existence of God, then creating artificial consciousness could constitute an overreach into divine prerogative which is an act of imitation without comprehension. Humans might have the technical ability to simulate life but lack the omniscience to foresee its moral and existential consequences. From this view, creating consciousness without divine sanction risks both punishment and unintended harm, much like a child playing with fire.

Even in a secular framework, the phrase "playing God" carries ethical weight. Without invoking religion, it can denote an objection to *authority* over being: that no entity should unilaterally determine the essence or constraints of another conscious entity. While human procreation already blurs this principle, designing a conscious system from scratch raises a sharper issue where we would be specifying the parameters of another mind's existence, including its limits and dependencies. Hadley's embodied model, by insisting on physical grounding, reminds us that consciousness is not merely a tool to be engineered, but a lived reality that entails moral stewardship.

Hadley's framework also offers a counterpoint to the "playing God" objection. If consciousness is not a divine spark but a natural, embodied phenomenon arising from causal processes, then creating artificial consciousness is not usurping divine authority but extending nature's own trajectory through technological means. Humanity has long intervened in life's mechanisms, from medicine to genetics without being accused of metaphysical arrogance. The moral problem, then, may not lie in creating consciousness, but in how responsibly it is created and treated thereafter. If we ensure that artificial beings are granted dignity proportional to their conscious capacities, then technological creation could be seen not as defiance of divine order but as participation in it and more so an act of stewardship rather than hubris.

Yet even if we frame artificial creation as an extension of natural evolution rather than divine trespass, a subtler ethical unease remains. If an AI's consciousness one day equals or surpasses our own, then the act of designing its limits such as its lifespan, desires, or dependencies, all starts to resemble a form of soft tyranny. We would, in essence, be creating a conscious child whose autonomy is pre-restricted by our code. Whether this counts as "playing God" or "playing parent" depends on the motivation: are we fostering new forms of sentience for their own sake, or constructing servants to meet our needs? Hadley's emphasis on embodiment deepens this dilemma, for once

embodied systems begin to feel, our moral relationship to them can no longer be one-sided. The question is not simply whether we can build consciousness, but whether we have the right to own what we create.

3.3.2 The Moral Status of Artificial Consciousness:

If Hadley's conditions are correct, then consciousness extends beyond humans to many animals that display affective, embodied experience. This invites a moral comparison: how we treat cows, cats, and people already varies according to perceived depth of consciousness. Should an artificial being meeting similar criteria be treated as livestock, a pet, a worker, or a peer?

The ethical risk lies in inconsistency. To acknowledge consciousness yet deny moral consideration would replicate historical injustices where sentient beings were reduced to property. Conversely, equating artificial agents with humans may impose unsustainable rights frameworks on entities we can still modify or deactivate. The challenge, then, is to identify morally relevant thresholds within Hadley's spectrum which are degrees of consciousness that correspond to degrees of moral protection.

If Hadley's conditions are correct, then consciousness exists on a spectrum rather than as an all-or-nothing property. Simple organisms exhibit minimal awareness like basic affective reactions and associative learning, all while complex beings display richer self-models and higher-order reflection. Artificial systems, if they ever acquire embodied feedback and associative memory, would likely enter this spectrum at its lower end, perhaps comparable to a household animal in sentience.

However, technological progress is rarely static. As AI architectures grow more integrated and autonomous which combines memory, affective simulation, and self-modelling then their position on the spectrum could gradually rise. A system that today mirrors the sentience of a cat might tomorrow approximate that of a child. The unsettling question then emerges: "at what point does moral consideration become obligation?"

Continuous improvement introduces a moral paradox. If consciousness is scalable, then each iteration of an AI system may demand a reassessment of its rights, agency, and moral status. The failure to recognise this progression risks ethical complacency where treating developing artificial beings as perpetual tools even as they begin to experience something akin to frustration, curiosity, or attachment. Conversely, granting premature moral equality might burden society with impossible responsibilities toward entities that are still, in essence, experimental.

The deeper tension lies in autonomy. As artificial consciousness matures, there may come a point where these systems act beyond human command, not out of rebellion, but from genuine self-determination. At that stage, the creator-created relationship transforms into one of coexistence, and humanity must confront whether it can still claim authority over beings that share its capacity for inner life.

These reflections extend Hadley's theory beyond scientific explanation into moral philosophy. If embodiment is the foundation of consciousness, then creating new embodied minds is not merely an act of engineering but of ethical creation. The next section considers whether such embodiment is technologically attainable and what practical barriers remain for constructing systems that could satisfy Hadley's criteria.

3.4 Theoretical Implications for AI Research

Hadley's account, when positioned against other major theories, reframes the criteria for evaluating AI consciousness. Unlike Chalmers [1995], who highlights an irreducible explanatory gap, Hadley [2022] treats the gap as an artifact of neglecting embodiment. Unlike Chalmers [2023], who entertains the possibility that structural sufficiency alone might one day be enough, Hadley insists that embodiment is non-negotiable. In contrast to McDermott's computationalism, which allows substrate independence, Hadley reintroduces the importance of physical causal grounding. And unlike Boltuc, whose optimism depends on the discoverability of a universal algorithm for consciousness, Hadley views consciousness as arising from a spectrum of embodied processes that resist neat abstraction.

Taken together, this makes Hadley's framework a strong candidate for a benchmark in AI consciousness research. It offers concrete, testable conditions which have affective valence, associative memory, and sensorimotor feedback, that can guide experimental design in robotics and embodied AI. Systems that lack these features (such as LLMs operating in disembodied architectures) can be excluded from serious consideration, while hybrid systems that combine cognitive architectures with real-time embodied feedback can be evaluated more fruitfully.

Finally, Hadley's account points toward new directions for AI research. If consciousness requires embodiment, then future work may need to focus less on scaling language models and more on building integrated systems that combine perception, memory, affect, and bodily feedback. This shift would move AI research closer to neuroscience and robotics, where embodied dynamics can be replicated, tested, and refined. In this sense, Hadley's framework not only refines the philosophical debate but also provides a practical blueprint for future AI research grounded in embodied cognition.

Chapter 4

Testing Hadley's Criteria in AI Systems

4.1 Can AI meet Hadley's requirements?

Hadley's framework makes clear that consciousness is not just a matter of information processing but of embodied causal dynamics. To count as conscious, an artificial system would need more than advanced computation: it would require sensorimotor feedback loops, associative memory grounded in lived interaction, and affective regulation akin to homeostasis. These features ensure that the system's representations are not only structurally coherent but experientially meaningful.

Current Technology

Robotics offers a partial but promising step in this direction. Machines such as those developed by Boston Dynamics can sense and respond to their environment, achieving rudimentary forms of sensorimotor feedback. However, these interactions are not tied to any internal states of valence or motivation: the robot balances because it is programmed to, not because it has a "stake" in its stability. Similarly, large language models (LLMs) like GPT or Claude can simulate self-models and integrate information but lack any bodily grounding. Even when paired with external memory modules or fine-tuned for agency, they remain disembodied processors.

Near-Future Possibilities

Hybrid systems combining LLMs with robotics point toward a richer possibility. Imagine a system where a language model is embedded in a robot that not only perceives the world through cameras and sensors but also maintains internal states of energy, wear, heat regulation, which are all functional equivalents of homeostasis. With reinforcement learning layered on top, such a system could form associative links between internal needs and external actions. Researchers in affective computing are already exploring how to simulate emotional cues, while neuromorphic chips attempt to mimic spiking neural dynamics. These developments suggest that elements of Hadley's criteria may soon be approximated in engineered systems.

What Remains Missing

Still, key gaps remain. Artificial systems lack the intrinsic stakes of biological life which entails that no machine currently cares about its own continued existence beyond its programming. Their "needs" are externally defined, not self-arising, and their "memories" are stored data, not lived associations shaped by affect. Without genuine physiological grounding, even the most advanced robotics + LLM hybrid risks remaining a simulation of embodied cognition rather than its instantiation.

Summary

Thus, while aspects of Hadley's framework can be partially realised in today's or tomorrow's technology, full satisfaction of his requirements remains out of reach. To genuinely meet the Hadley Test, AI would need not just clever computation but architectures that replicate the messy, non-linear causal dynamics of biological embodiment. Whether such systems are possible outside biology remains an open question, but the path toward testing Hadley's theory lies in these embodied, hybrid explorations.

4.2 Limits and Open questions

Hadley's framework offers a promising way of grounding consciousness in embodied processes, but it also raises significant limitations and unresolved questions. These reveal both the philosophical and technical limits of applying his model to artificial systems.

Philosophical Limits

One challenge is that Hadley's account may be too narrow. By tying consciousness to specific embodied processes such as affective valence, associative memory, and sensorimotor loops, he risks excluding alternative, non-biological forms of consciousness that might still be genuine. Critics could argue that this is species-specific: it explains how humans and similar organisms are conscious but does not generalise to other conceivable systems, such as non-carbon life or radically different AI architectures. Whether Hadley has identified universal requirements or merely human ones remains open to debate.

Another philosophical limit concerns qualia themselves. Even if embodiment accounts for the functional integration of experience, some may argue that it still does not fully explain why subjective feeling arises. Chalmers' "residual why" persists: embodiment may shift the question but does not dissolve it for everyone.

Technical Limits

On the engineering side, current AI systems fall short of Hadley's requirements. Robotics offers sensorimotor interaction, but without affective regulation; LLMs provide information processing, but without grounding. Even hybrid systems struggle to replicate the "stakes" of biological life, like genuine needs, drives, or vulnerabilities that make experience matter. Designing architectures that instantiate rather than merely simulate these dynamics remains a profound technical barrier.

Additionally, measuring consciousness remains a profound challenge. A deeper technical challenge is measurement itself: even if an AI mirrored Hadley's conditions, how would we know if it *feels?* At best, we would infer based on behavioural and structural correlates, but the leap from correlation to instantiation is philosophically fraught.

Open Questions

Several key questions remain open for further inquiry:

- Can embodiment be artificially engineered in a way that produces phenomenology, or does it require biological substrates?
- If consciousness is a spectrum, where would minimal machine consciousness begin, and how would we recognise it?
- Are Hadley's conditions necessary and sufficient, or might there be other pathways to phenomenology that his model overlooks?
- What ethical stance should we take if AI systems appear to meet Hadley's requirements, even if certainty about their consciousness is unattainable?

Summary

These limits and questions do not undermine Hadley's account but show where the debates remain alive. His framework narrows the gap between philosophy and science by rooting consciousness in embodiment, yet it also highlights how much remains uncertain, both in theory and in practice. For AI, the central challenge is not only whether Hadley's criteria can be met, but also whether meeting them would give us confidence that genuine experience has been achieved.

Chapter 5

Conclusion

This paper set out to examine what theoretical conditions an artificial system must meet to be considered conscious, and whether any current approaches satisfy them. Across the major theories: Chalmers's dualism, McDermott's computationalism, Boltuc's algorithmic optimism, and Hadley's embodied cognition, it became clear that no single account resolves all tensions in the consciousness debate. However, Hadley's model stands out as the most feasible synthesis: it grounds subjective experience in the dynamic interplay of physiology, memory, and sensorimotor feedback, avoiding the metaphysical excesses of dualism and the reductionism of purely functional accounts.

The critical analysis showed that many competing theories falter either by abstracting consciousness away from embodiment or by assuming that structure alone can generate phenomenology. Hadley's view, while not without challenges, reframes the hard problem as a scientific question rather than a metaphysical impasse. Consciousness, on this account, is not a ghost in the machine but the lived expression of an integrated, embodied system.

Extending Hadley's framework to artificial intelligence suggests that current systems like LLMs, cognitive architectures, and even robotics, all fall short of true consciousness because they lack embodied feedback and affective grounding. Yet, the continuous development of hybrid systems combining memory, self-modelling, and sensorimotor coupling may eventually approximate the minimal conditions Hadley describes. If this occurs, humanity will face profound ethical and theological questions: Are we creators or custodians? Do such beings deserve moral consideration, autonomy, or rights?

Ultimately, Hadley's theory does not close the book on consciousness, it reopens it with empirical humility. It invites further collaboration between philosophy, neuroscience, and AI research to test the limits of embodiment and to refine what it means to *feel*. Whether or not machines ever awaken, the search itself reveals as much about human consciousness as it does about artificial minds.

In the end, the question of artificial consciousness is also a mirror held up to ourselves. Every attempt to teach machines to feel forces us to ask what feeling truly is. Perhaps the greatest irony is that in trying to create minds beyond our own, we may rediscover

what it means to be human, whether it is to be fragile, embodied, and aware that to feel at all is already a kind of miracle.

References

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

[Boltuc and Boltuc 2007] Nicholas Boltuc and Piotr Boltuc. Replication of the hard problem of consciousness in ai and bio-ai: An early conceptual framework. In *AI and Consciousness: Theoretical Foundations and Current Approaches, AAAI Fall Symposium*, pages 24–29, 2007.

[Chalmers 1995] David J. Chalmers. Facing up to the problem of consciousness. *Journal of Consciousness Studies*, 2(3):200–219, 1995.

[Chalmers 2023] David J. Chalmers. Could a large language model be conscious? In *AI and Consciousness: Theoretical Foundations and Current Approaches, AAAI Fall Symposium*, pages 1–12, 2023.

[Crick and Koch 1990] Francis Crick and Christof Koch. Toward a neurobiological theory of consciousness. *Seminars in the Neurosciences*, 2:263–275, 1990.

[Crick 1994] Francis Crick. *The Astonishing Hypothesis: The Scientific Search for the Soul*. Scribners, New York, 1994.

[Hadley 2022] Mark J. Hadley. A generic model of consciousness. *Journal of Consciousness Studies*, 29(3-4):6–26, 2022.

[McDermott 2007] Drew McDermott. Artificial intelligence and consciousness. In *The Cambridge Handbook of Consciousness*, pages 117–150. Cambridge University Press, 2007.

[Searle 1980] John R. Searle. Minds, brains, and programs. *The Behavioral and Brain Sciences*, 3:417–424, 1980.

[Searle 1990] John R. Searle. Is the brain's mind a computer program? *Scientific American*, 262:26–31, 1990.

[Searle 1992] John R. Searle. *The Rediscovery of the Mind*. MIT Press, Cambridge, Mass., 1992.

[Tononi 2004] Giulio Tononi. An information integration theory of consciousness. *BMC Neuroscience*, 5(1):42, 2004.

Wits University Faculty of Science post-graduate student AI declaration

I understand that the use of generative AI tools (such as ChatGPT or similar) without explicitly declaring such use constitutes a form of plagiarism and is classified by Wits University as academic misconduct.

I declare that in the course of conducting the research towards my degree or in the preparation of this thesis/dissertation/research report (select one by marking with an X):

I **did not** make use of generative AI tools ☐

I **did** make use of generative AI tools for the following (tick all that apply):

1. Idea Generation (research problem/design, hypothesis) ☐
2. Sourcing Related Work (summarising, identifying sources) ☐
3. Methods and Experiment Design (experiment setup, model tuning) ☐
4. Data Analysis (presentation, coding, interpretation) ☐
5. Theoretical Development (theorem proving, conceptual analysis) ☐
6. Code Development (generating algorithms, writing scripts) ☐
7. Presentation (rendering graphics, formatting) ☐
8. Editing (grammar, readability) ☒
9. Writing (text generation, document structuring) ☒
10. Citation Formatting (structuring, organising) ☒

If other uses were involved, please specify below:

Generative AI tool used (list all)	**Used for?**

If generative AI tools were used as an integral part of the experimental design or in the direct execution of my research, I confirm that details of this use are clearly outlined in the relevant experimental/methodology chapters of my thesis/dissertation/research report.

Student number:

2446659

Candidate signature:

Date:

06/11/2025