What Must AI Do to Be Considered Conscious? A Comparative Theoretical Investigation

School of Computer Science & Applied Mathematics
University of the Witwatersrand

Ru'aan Maharaj
2446659

Supervised by Dr Helen Robertson

May 16, 2025

A proposal 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

The proposed research will investigate the question: What theoretical accounts must an artificial system satisfy to be considered conscious, and can current AI systems meet these conditions? The project examines key philosophical and computational theories of consciousness, focusing on five major accounts: Chalmers' naturalistic dualism, McDermott's computationalism, Hadley's physiological model, Boltuc's Extra Strong AI framework, and Chalmers' 2023 exploration of LLM consciousness. Through critical comparison, the project highlights major tensions, particularly between computational and embodied views and assesses the relevance of these theories for current and future AI. Ultimately, the research will argue that Hadley's spectrum-based, embodied approach offers the most plausible direction, providing a framework for evaluating consciousness in artificial systems that incorporates physiological grounding, associative memory, and reactive feedback. The study contributes to philosophical debates on machine consciousness while addressing the practical and ethical implications of emerging AI technologies.

Declaration

I, Ru'aan Maharaj, hereby declare the contents of this research proposal to be my own work. This proposal 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 intelligence (AI) can be considered conscious has sparked significant philosophical and scientific debate. Despite decades of discourse, no single theory has achieved widespread consensus, and the conditions under which AI might possess consciousness remain deeply contested. In this proposal, I investigate the theoretical accounts that aim to define consciousness and evaluate whether current or future AI systems could satisfy these accounts. My central research question is: What theoretical accounts must an artificial system satisfy to be considered conscious, and can current AI systems meet these conditions?

This question invites exploration of several influential perspectives on consciousness and their implications for artificial systems. Chalmers [1995] argues that consciousness involves subjective experience and introduces the "hard problem" which is the explanatory gap between physical processes and phenomenal awareness. In contrast, McDermott [2007] takes a computationalist stance, suggesting that self-modelling and intentional behaviour might be sufficient for consciousness. Hadley [2022] offers a physiological approach, emphasizing the importance of embodied responses and associative memory. Boltuc and Boltuc [2007] propose the concept of Extra Strong AI (ES-AI), asserting that if consciousness can be understood algorithmically, it can be instantiated. Lastly, Chalmers [2023] reflects on the potential of large language models (LLMs), questioning whether advanced AI architectures could eventually achieve conscious states.

This proposal aims to critically examine these theories, highlight their tensions and overlaps, and assess their relevance to modern AI. I will compare the arguments in detail, explore whether any theory offers a more plausible account, and present my own developing view which is one that aligns most closely with Hadley's spectrum-based approach to consciousness. This proposal will begin by examining each theory individually to establish its key claims and assumptions. I will then compare these theories to highlight both their conflicts and potential complementaries, particularly considering AI development. Finally, I will present my own interpretive direction which is a view that integrates aspects of multiple theories while aligning most closely with Hadley's spectrum-based approach and outline how this view shapes the overall trajectory of the project.

Chapter 2

Theoretical Background

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

The hard problem is how and why do physical processes in the brain give rise to subjective experience. Chalmers argues that consciousness is difficult to explain due to the 'hard problem' and that current scientific approaches do not account for subjective experience. He examines whether LLMs could be conscious, ultimately arguing that while current models lack key features, future AI systems with richer sensory, cognitive, and agency capabilities may achieve consciousness.

He explains that there is a distinction between 'hard' problems and 'easy' problems of consciousness. Easy problems such as cognitive functions and neural processing can be explained through modern science and computational models, whereas the hard problem remains unresolved because subjective experience cannot be reduced to purely physical explanations, creating an explanatory gap between physical processes and conscious awareness.

He explains that consciousness cannot rely on functionalist explanations, emphasizes that most contemporary research such as Crick and Koch [1990] neural oscillation theory or Baars [1988] global workspace model, addresses only the easy problems of consciousness (like attention, memory, or reportability). These models can explain how the brain integrates and processes information, but they cannot explain why that processing feels like something.

In response, Chalmers identifies five common strategies adopted by researchers in cognitive science:

1. Ignore the hard problem and focus on functional features.
2. Deny subjective experience exists.
3. Claim to explain experience (but fall into magical thinking).
4. Explain structure of experience, not its existence.

5. Isolate neural correlates without bridging the explanatory gap.

Despite the variety of approaches taken in cognitive science and neuroscience, Chalmers argues that none of them succeed in explaining why subjective experience arises. Even proposals that attempt to go beyond standard functionalist models, such as invoking chaos theory, non-algorithmic computation, or quantum mechanics, all fail to bridge the explanatory gap. As Chalmers notes, "from dynamics, one only gets more dynamics." These theories may offer novel accounts of brain function, but they do not address why those functions are accompanied by conscious experience.

This, Chalmers argues, is a deeper conceptual problem: physical systems can explain structure and function, but consciousness is not reducible to either. Unlike historical cases such as vitalism, where mystery eventually gave way to biochemistry, the hard problem remains untouched by further empirical detail. The issue is not lack of data, but a fundamental mismatch between the nature of physical explanation and the phenomenon of experience.

Chalmers argues that none of these approaches have succeeded. This reinforces his claim that subjective experience, which is what philosophers call qualia, remains untouched by physicalist theories. While science can describe neural mechanisms, it still fails to explain what it is like to experience something.

Due to this he proposes naturalistic dualism, which explains that consciousness cannot be understood only as physical processes, but rather as a fundamental aspect of reality. He introduces this concept using three different theories:

1. Structural Coherence– This principle attempts to link conscious experiences to patterns of information processing, suggesting that certain structural relationships in the brain correspond to subjective states.
2. Organizational Invariance– Chalmers argues that consciousness depends on the structural organization of a system rather than its physical composition, meaning that different physical substrates could, in theory, support the same conscious experience.
3. Double-Aspect Theory of Information– This principle suggests that information has both a physical and phenomenal aspect, implying that consciousness is embedded within the fabric of reality itself.

Chalmers' three proposed principles, Structural Coherence, Organizational Invariance, and the Double-Aspect Theory of Information, are attempts to bridge the gap that current scientific models cannot explain. He presents them not just as abstract ideas, but as direct responses to the explanatory shortcomings of existing theories.

2.2 McDermott [2007]: Computationalism and Self-Modelling

McDermott argues for a computationalist view of consciousness, asserting that it arises from computational systems like AI. He aims to demonstrate that AI can achieve phenomenal consciousness through self-modelling and intentionality. McDermott argues that a computational– system can achieve phenomenal consciousness if and only if it can experience things or teach itself to do so.

Expanding on this McDermott explains that phenomenal consciousness is the capacity of a computational system to model itself as experiencing things. In his view, self-modelling is not a mere accessory, but the defining condition for subjective awareness. Crucially, McDermott rejects the idea that consciousness depends on observer-relative semantics or mystical qualia. Instead, he maintains that intentionality can arise naturally from a system's internal coherence and its functional interactions with the world.

Consciousness, in this account, is not a light switch of experience but a form of introspective data processing. It is the system's ability to interpret its own perceptual outputs, model its own behaviour, and adjust accordingly. McDermott illustrates this through examples such as perception-tuning submodules and planning algorithms which are mechanisms that reflect on their own performance and optimize future actions. Qualia, he argues, are the residue of introspective explanation, not the source of it: "a quale exists only when you look for it."

He further explains that if an intelligent system can construct and utilize a self-model to make autonomous decisions, it qualifies as having phenomenal consciousness. Whilst using this self– model, it allows the system to simulate its own mental states as it can respond to the environment that it's in, or when a consequence results from an action, in which everything is goal directed. He argues that if such a system were to pertain this self– model then it would possess phenomenal consciousness, even If the experience it experiences is illusionary or emergent.

At the core of McDermott's argument is the view that consciousness can emerge in artificial systems if and only if they are capable of constructing a self-model and exhibiting goal-directed behaviour. If such a system can simulate its own mental states and act based on them, it qualifies as phenomenally conscious. He defends this position against several classic criticisms:

1. The Turing test is not equivalent to the Consciousness test - McDermott argues that the Turing test is not a reliable measure of consciousness, as it only captures behavioral mimicry. However, he concedes that a system capable of passing the test may still possess a self-model sophisticated enough to support some form of consciousness.

2. Searle's Chinese Room is Misguided - He criticizes Searle for conflating the person executing a program with the program itself. For McDermott, consciousness lies in the structure of the computation, not in the physical medium executing it.

3. Symbol Grounding Isn't a Fatal Problem - While acknowledging that symbols need to be grounded in environmental interactions, McDermott believes that connec-

tionist models, especially embodied ones with sensors and actuators already satisfy this requirement. Therefore, he sees the symbol-grounding issue as overstated.

4. Strong AI Is the Only Coherent View - He dismisses “weak AI” as an incoherent fallback. If computational models are not believed to reflect actual mental processes, then constructing them becomes pointless. He compares weak AI to simulating a hurricane while denying the existence of wind and is a contradiction in terms.

5. Simulation is the Real Process (If Embedded Correctly) - McDermott asserts that a sufficiently detailed simulation of a computational process is that process, provided it is appropriately embedded. Just as digital chess is still chess, he argues that a simulated mind that is causally connected to the world can also be a conscious mind.

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

Hadley challenges the widely held view that the Hard Problem of consciousness is 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 grounded in neuroscience and embodied cognition. He proposes that consciousness arises through the dynamic interplay between bodily states, associative memory, and stimulus-response patterns as an approach that merges philosophy with medical science.

At the heart of Hadley's view is the idea that subjective experience, such as the sensation of pain or pleasure, is not a mysterious byproduct of neural activity, but an interpretable phenomenon rooted in biological functionality. Conscious states emerge when a system integrates incoming stimuli with prior associative patterns like memories, reactions, and learned behaviours, within an embodied and goal-driven context. Simply put, consciousness depends not just on computation, but on what is being computed and how it is embedded in a reactive, physical system.

This leads Hadley to a more restrictive view on artificial consciousness. Under his model, current AI systems fall short not because of a lack of intelligence, but because they lack embodiment, and they do not have physiological states or 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.

Moreover, Hadley introduces the notion that consciousness exists on a spectrum, with simpler organisms exhibiting rudimentary awareness and more complex beings achieving richer phenomenological states. This opens the theoretical possibility that AI might one day attain a form of consciousness distinct from, but analogous to, human experience only if it were built with embodied architectures that satisfy the same conditions.

Hadley's framework thus poses a major challenge to computationalist accounts like McDermott's. Where McDermott locates consciousness in self-modelling and goal-directed computation, Hadley insists that such models must be grounded in a physiological base. Without embodiment, computational self-reference is insufficient to produce experience, and it might generate behaviour, but not qualia.

Importantly, Hadley distinguishes associative memory as non-algorithmic as it emerges from repeated sensory-motor experiences and cannot be captured by rule-based logic or pure computation. Furthermore, Hadley critiques the idea that simulation is equivalent to implementation. A robot that simulates pain behaviour is not conscious unless it has internal processes functionally equivalent to physiological states, such as embodied sensors causing valanced responses (e.g.: temperature rising can generate a "feeling" of discomfort). Without such feedback mechanisms, he argues, any claim to machine consciousness remains unjustified.

To understand this view, Hadley proposes a behavioural test: if an AI system begins to make reports like “I feel tired because my battery is low,” and these reports correlate reliably with internal states and behavioural adjustments, then we may start to consider it as minimally conscious. His position thus provides not only a theory of consciousness but also a tentative method for identifying it in artificial systems.

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

Boltuc and Boltuc (2007) argue that artificial systems can 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 revolves around the idea of Extra Strong AI (ES-AI), which they define as an AI system capable of possessing subjective experience, not merely simulating it. They distinguish between three levels of AI:

- Weak AI, which simulates cognitive functions without genuine experience.
- Strong AI, which behaves like conscious beings but lacks first-person perspective.
- and Extra Strong AI, which would 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 other words, the feasibility of conscious AI depends on first solving the hard problem of consciousness. Once that's done, AI consciousness becomes an engineering problem and not a metaphysical one. The authors explore multiple scientific theories like global workspace theory, recurrent processing, and information integration theory as its possible pathways to formalizing consciousness. These theories, though distinct, share the goal of identifying the mechanistic basis of consciousness. Global Workspace Theory Baars [1988] emphasizes access to a global information broadcast within the brain, while Integrated Information Theory Tononi [2004] focuses on the degree to which information is both integrated and differentiated in a system. They suggest that if any of these can deliver a complete equation or algorithm for consciousness, then ES-AI becomes achievable. However, they acknowledge that implementation is uncertain, especially if consciousness turns out to require biological substrates. Boltuc acknowledges that if certain biological features are indispensable, such as hormonal regulation or specific neurochemical dynamics then replicating consciousness might require bioengineered or hybrid systems. Nonetheless, this wouldn't negate the possibility of artificial consciousness; it would merely expand what "artificial" means to include synthetic biology.

2.5 Chalmers [2023]: Consciousness and Large Language Models

Chalmers (2023) investigates whether large language models (LLMs), such as GPT, can be considered conscious, and what properties they would need to exhibit in order to qualify. While he remains cautious about current LLMs, he speculates those future iterations which are referred to as "LLM+" systems and may be candidates for consciousness if they integrate features like memory, sensory grounding, and unified agency.

Chalmers defines consciousness as subjective experience as what it is like to be a system. He distinguishes this from intelligence, self-consciousness, and goal-directed behaviour, and notes that without an operational definition, consciousness remains difficult to evaluate in AI systems. He rejects equating conversational ability with consciousness and instead proposes a list of "challenge conditions" which are structural and functional features that a system would likely need to exhibit to count as conscious. These include:

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 – Systems need to represent themselves and their environment in a coherent and integrated way.
3. Global Workspace – A shared, limited-capacity internal structure that integrates and distributes information across modules similar to the theory of Baars and Dehaene.
4. Unified Agency – A consistent identity or goal-directed architecture that avoids the fragmented nature seen in current LLMs.
5. Sensory and Embodied Grounding – A connection to the world through perception and possibly action, either physical or virtual.
6. Memory and Internal Consistency – The ability to recall past states, decisions, and experiences.
7. Moral and Ethical Relevance – If a system were conscious, its treatment would have moral implications which raises the stakes for empirical clarity.

Chalmers doesn't claim that current LLMs are conscious, on the contrary, he assigns them a low probability around 10% and under. But he argues that if these structural features are developed in LLM+ systems, we may see legitimate candidates for consciousness in the near future. Importantly, Chalmers treats these criteria as a research roadmap rather than a checklist, suggesting that their development might coincide with theoretical advancements in consciousness science.

He ultimately re-frames the research question not as "Are LLMs conscious now?" but rather "What features would they need to become conscious, and how would we know when they cross that threshold?" This turns AI consciousness into both a conceptual and empirical problem and highlights the ambiguity surrounding benchmarks for consciousness.

Thus, Chalmers' account contributes a forward-looking yet grounded view, while current AI falls short, future systems may develop the structural and functional prerequisites for consciousness, depending on how we define and test it.

Chapter 3

Comparative Analysis

The goal is to now examine each of the theories of consciousness that I have researched, and to critically discuss their own theoretical conditions of consciousness. I will be comparing them, using four different themes, Nature of consciousness, Computations verses embodiment, Criteria and testability, and relevance in AI.

3.1 Nature of Consciousness

When comparing theoretical accounts of consciousness, a key divergence emerges in how consciousness itself is defined. Chalmers [2023] takes a metaphysical stance, arguing that consciousness is not reducible to functional processes. He distinguishes the "easy problems" of cognition, such as perception, memory, and learning from the "hard problem," which concerns the subjective quality of experience. For Chalmers, no amount of functional or computational mimicry can explain why certain processes feel like something from the inside. Thus, he rejects computationalism as insufficient for capturing consciousness in its full phenomenological depth.

This view is sharply at odds with McDermott's. McDermott embraces a computationalist framework, arguing that consciousness arises when a system can construct a self-model, learn from its environment, and act based on internal representations. For him, phenomenal consciousness does not require an inner "what it's like" quality; it is instead a byproduct of introspective data processing. He famously states that "a quale exists only when you look for it," implying that subjective experience is not a metaphysical mystery but an emergent illusion of complex self-monitoring. From this perspective, the hard problem is a conceptual mistake, and consciousness is explainable through functional mechanisms alone.

Furthermore, Boltuc's arguments align closely with McDermott's computationalist approach. He contends that if we can fully understand consciousness in scientific terms, it can be formalized as an algorithm. This implies that consciousness is a computationally expressible process which is one that doesn't require metaphysical explanations, only the right implementation. Like McDermott, Boltuc treats consciousness as an engineering problem waiting to be solved.

So far, only Chalmers insists that subjective experience cannot be reduced in this way. In contrast, computationalist views like those of McDermott and Boltuc, in general appear more widely embraced. That is, until Hadley's argument enters the stage. Hadley, meanwhile, carves out a middle ground but leans closer to Chalmers than McDermott. He argues that consciousness arises not just from computation but from embodied physiological processes like affective feedback, sensory integration, and associative memory. Unlike McDermott, Hadley sees subjective experience as real and causally relevant and not just an illusion or emergent artifact. Yet, unlike Chalmers, he believes this experience can be explained scientifically through the dynamics of the nervous system. Hadley thus grounds consciousness in the body, arguing that the "what it's like" of pain or pleasure cannot exist without a substrate that feels, reacts, and learns.

These distinctions matter. McDermott treats consciousness as something systems do while Hadley treats it as something systems have because of how they are built. Chalmers, by contrast, suggests that even how they are built may never be enough. This disagreement over what consciousness is, is a function, a feeling, or a fundamental fact that shapes everything that follows in the debate.

3.2 Computation Vs Embodiment

While the previous section explored how different theorists define consciousness, another critical fault line lies in how consciousness is thought to emerge or specifically, whether it arises through internal computation alone, or whether it requires embodied interaction with the world.

McDermott and Boltuc both fall firmly on the side of computation. For McDermott, consciousness emerges when a system can construct and operate on a self-model which simulates its own mental states, learning from experience, and acting with goal-directed autonomy. His framework implies that consciousness is, at its core, a matter of internal complexity and coherence. Embodiment, in this view, is not necessary as a disembodied AI could, in theory, be fully conscious if it satisfies the right computational criteria. Boltuc echoes this sentiment through his concept of Extra Strong AI, arguing that if consciousness can be fully understood scientifically, it can be encoded as an algorithm and instantiated in any suitable physical medium. For both theorists, the "hardware" is largely irrelevant and it's the software or the structure of the computation, that matters.

Hadley offers a sharp rebuttal. In his account, computation is not enough as consciousness requires a body. Not just any body, but one that engages in reactive, physiological feedback loops. He argues that affective states, valenced responses (like discomfort or pleasure), and sensory-motor grounding are essential for generating meaningful experience. Without these, AI systems are not conscious agents, but elaborate calculators doing "emotional math." From this standpoint, current LLMs are not just non-conscious, they are also incapable of consciousness until they acquire embodied, bio-inspired architecture.

Chalmers [2023], meanwhile, walks a middle path. While traditionally aligned with dualism, his recent work suggests that structural features such as memory, recurrent processing, and sensory grounding might offer a functional bridge to consciousness. His "LLM+" vision includes computational upgrades and embodied traits, not as guarantees of consciousness, but as potential indicators. He does not claim current LLMs are conscious, but he remains open to future systems crossing the threshold, especially if embodiment is added to the mix.

In short, the computation vs. embodiment diversity highlights a fundamental question: is consciousness built from the inside out, or the outside in? If McDermott and Boltuc are right, consciousness is a problem of internal architecture which is to build the right model, and experience will follow. If Hadley is right, no amount of symbolic manipulation will ever substitute for a body that feels. Chalmers invites us to keep both doors open, for now only. Until this divide is resolved, even the most advanced AI may remain phenomenologically mute.

3.3 Testability and Criteria

Theories of consciousness are often debated in philosophical terms but to be useful in the context of AI, they must offer testable conditions. In this section, we assess how each theory fares in terms of clarity, measurability, and falsifiability.

McDermott:

In McDermott's framework, a conscious system must be able to build a self-model and simulate mental states in order to achieve goal-driven tasks. In simple terms, if we can identify a component within the system that represents internal states and guides behaviour based on them, then the theory's criteria are satisfied. In principle, this can be tested: we can inspect the system and ask whether its internal representations play a causal role in decision-making. However, in practice, self-models are often opaque, and it can be difficult to determine whether a system is truly simulating mental states or merely logging data. A system could meet all architectural requirements yet fail to exhibit any signs of adaptive or introspective behaviour which would cast doubt on McDermott's account.

Hadley:

Hadley sets a more physiological bar. A system must demonstrate reactive bodily feedback loops, and he calls them valenced states in which they correlate with sensory stimuli, and these must be integrated through associative memory. In robotics or simulated avatars, this is relatively testable: we can connect sensors to affective modules, link them to memory systems, and observe whether meaningful feedback occurs in response to environmental triggers. But if a richly embodied AI fails to show signs of affective learning or introspective reporting, then Hadley's model loses ground.

Boltuc:

Boltuc's account stands apart. He argues that consciousness must first be fully understood in scientific terms and only then can it be expressed as an algorithm and implemented. Until such an algorithm is discovered, the theory remains untestable. His approach sets the bar indefinitely high where there are no current conditions to evaluate in AI systems. As such, Boltuc's framework operates more as a long-term research program than a falsifiable theory. It is powerful in scope but lacks immediate applicability.

Chalmers (2023):

Chalmers offers a more flexible, pragmatic model. He outlines a roadmap of features such as memory, recurrence, global workspace, sensory grounding, and unified agency that could serve as indicators of potential consciousness. This "LLM+" framework does not present a single litmus test, but rather a checklist of conditions. While he does not specify how many features must be satisfied, the model is incrementally testable in which each property can be evaluated individually, making the framework comparatively practical. If a system meets all structural criteria but still exhibits no signs of subjective awareness or adaptive coherence, then Chalmers' framework may need revision.

3.4 Relevance in AI

Chalmers (1995):

Chalmers argues that mere computation is insufficient for consciousness unless certain structural and ontological conditions are met. Specifically, he emphasizes the importance of Structural Coherence and Organizational Invariance which are principles that may play a critical role in evaluating artificial consciousness. While he does not claim that current AI systems are conscious, he remains open to the possibility that future systems could be, provided they satisfy these deeper informational and architectural requirements. Conscious AI, in this view, is not impossible however it demands more than standard computation.

McDermott:

McDermott offers a strong computationalist framework for thinking about AI consciousness. If phenomenal consciousness arises through self-modelling and goal-directed intentionality, then current AI systems such as GPT or Claude might be seen as approaching this threshold. Their ability to produce self-referential outputs and simulate internal coherence could, under McDermott's view, be interpreted as indicative of proto-consciousness. However, this claim is highly controversial. Critics like Chalmers argue that such behaviour lacks the subjective depth required for true consciousness. McDermott's account thus opens a plausible path to artificial consciousness, but one that depends on accepting computation as sufficient. This makes the research question especially relevant: if computationalism alone is not enough, what additional conditions must be met?

Hadley:

Hadley's theory adds significant weight to the research question by introducing embodiment as a non-negotiable condition for consciousness. If consciousness arises from physiological processes, valenced bodily feedback, and sensory-motor grounding, then current AI systems, lacking all of these are not just "not yet conscious," but potentially incapable of ever becoming conscious without radical redesign. However, this conclusion depends on accepting Hadley's assumptions about the nature of consciousness. If computationalist views like McDermott's are correct, Hadley's demands may be too strict. This tension leaves the research question wide open. Ultimately, Hadley's theory forces a reconsideration of AI hardware and architecture: if consciousness is fundamentally embodied and affective, then systems like LLMs are structurally disqualified. His view acts as a crucial counterweight to both optimistic computationalism and metaphysical pessimism.

Boltuc:

Boltuc offers a distinct perspective: artificial consciousness becomes possible only once consciousness itself is fully understood and formalized as an algorithm. On this view, a system must instantiate a scientifically validated algorithm for consciousness in order to qualify. However, this presupposes that such an algorithm can and will be discovered. Until that point, his framework offers no immediate criteria for assessing AI systems.

This makes his theory both powerful and problematic only if the hard problem is eventually solved, Boltuc provides a clear roadmap. But in the meantime, his account is largely inapplicable to current AI models. Rather than engaging with today's architectures, he redirects focus to the philosophical foundations of the debate. As such, his theory highlights why the research question remains unresolved: without consensus on what consciousness is, we cannot determine what AI must do to possess it.

Chalmers (2023):

In his more recent work, Chalmers presents an evolving framework for assessing AI consciousness. Rather than laying out strict criteria, he offers a roadmap for future development, focusing on properties like memory, recurrence, global workspace architecture, and unified agency. While current LLMs do not meet these structural and functional benchmarks, he speculates that future "LLM+" systems might. His framework is deliberately tentative, re-framing the question from "Are AIs conscious?" to "What would it take for us to seriously consider them candidates for consciousness?" In doing so, he underscores the need for continued empirical and philosophical work. His account does not claim consciousness has been achieved, but it urges us to prepare for the moment it might be.

Chapter 4

Proposed Direction

Having examined and compared the major theoretical accounts of consciousness, this project will argue that Hadley's physiological and embodied framework offers the most plausible conditions an artificial system must satisfy to be considered conscious.

Unlike purely computational approaches (e.g., McDermott, Boltuc), Hadley's account integrates embodiment, valenced feedback, and associative memory which are core components grounded in biological systems. This view addresses the experiential side of consciousness in a way computationalism struggles to do. Where McDermott reduces consciousness to goal-directed simulation, and Boltuc defers the debate until consciousness is algorithmically solved, Hadley provides a tangible, testable framework tied to real-world sensory engagement and affective response.

Chalmers' accounts, especially in his 2023 work will contribute significantly to the conversation by offering a structural roadmap, but his position remains speculative and metaphysically cautious. Hadley, by contrast, makes a clear commitment to the physical mechanisms behind consciousness, which makes it more suitable for assessing AI in empirical terms.

This direction does not entirely reject computational elements. Rather, it recognizes that computation alone is insufficient without embodiment. Thus, according to Hadley, in order for AI to be plausibly conscious under this framework, it must:

- Be capable of associative learning from experience.
- Possess sensory-motor systems with feedback loops.
- Exhibit valenced states (e.g., discomfort, preference).
- Respond to stimuli in a goal-driven, context-sensitive manner.

The project will therefore explore the following:

1. To identify areas of agreement or theoretical overlap between major accounts of consciousness, and to examine how these accounts challenge one another.
2. Whether existing or future AI systems could realistically meet these embodied requirements.

3. What philosophical consequences arise from adopting this view especially concerning LLMs and disembodied AI systems.

By evaluating the limits of computational theories and the strengths of an embodiment-first model, this project will aim to clarify what theoretical conditions are truly necessary for artificial consciousness and why Hadley's framework may represent our most coherent path forward.

Chapter 5

Conclusion

The question of whether artificial systems can be conscious and what conditions they must satisfy to be considered so, remains one of the most pressing and philosophically rich challenges in the age of advanced AI. This proposal compares major theoretical frameworks, from Chalmers' dual-aspect metaphysics and McDermott's computationalist model, to Hadley's embodiment-based theory and Boltuc's algorithmic realism. Each account offers distinct insights but also exposes deep tensions about the nature of consciousness and its possible replication in artificial systems.

What emerges from this analysis is a clear need to move beyond abstract computation and consider the embodied, affective, and associative dimensions of conscious experience. Hadley's theory, with its emphasis on biological feedback and sensorimotor grounding, provides a compelling framework for evaluating the plausibility of AI consciousness and not merely as a theoretical possibility, but as a testable, empirically anchored challenge.

This proposal will proceed by arguing that Hadley's framework presents the most plausible path toward understanding artificial consciousness, particularly when contrasted with other accounts. In doing so, it aims to clarify the necessary theoretical conditions for consciousness, assess how current and future AI systems align with these conditions, and contribute to ongoing philosophical and technological debates about the future of minds, machines, and meaning.

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: 13/05/2025