Beyond the Hard Problem: Integrating Philosophy, Neuroscience, and AI in Consciousness Research

Dr. Mumun Das

S. S. J. Mahavidyalaya, Affiliated to Utkal University, Bhubaneswar, Odisha, India

Abstract: ***This article synthesises recent advances in consciousness studies by integrating contemporary philosophical debates, neuroscientific findings, and emerging discourse on artificial intelligence. It critically addresses philosophical issues, including qualia, the Hard problem, panpsychism, and illusionism, in juxtaposition with empirical neuroscientific theories such as Global Workspace Theory (GWT) and Integrated Information Theory (IIT). Moreover, it examines how interdisciplinary approaches have enhanced theoretical clarity and increased empirical testability. Ultimately, it explores the provocative possibility of machine consciousness, examining whether artificial systems can achieve conscious states, offer novel perspectives and bridge diverse theoretical frameworks to advance the study of consciousness.***

Keywords: Consciousness, Qualia, Hard Problem, Panpsychism, Illusionism, Neural Correlates of Consciousness, Global Workspace Theory, Integrated Information Theory, Artificial Intelligence, Neurophilosophy, Machine Consciousness.

1. Introduction

The study of consciousness remains one of the most profound and enduring mysteries in intellectual inquiry, continuously captivating philosophers and scientists alike. Central to this field is the "hard problem," famously articulated by the philosopher David Chalmers, which addresses why and how subjective experiences, or qualia, arise from physical processes within the brain (Chalmers, 1995). Philosophical discourses have extensively grappled with the intricate nature and ontological reality of these subjective experiences, sparking persistent debates between competing explanatory frameworks. Traditional physicalist approaches, which aim to explain consciousness solely through physical processes, have faced substantial criticism for their perceived inadequacy in addressing the subjective character of consciousness. In response, alternative theories, such as panpsychism and illusionism, have emerged, revitalising the philosophical landscape. Panpsychism posits that consciousness is a fundamental and pervasive aspect of reality, suggesting that even the simplest physical entities possess rudimentary experiential properties (Goff, 2019). Conversely, illusionism argues that our intuitive understanding of subjective experience is misleading, asserting that qualia may merely be cognitive illusions generated by complex neural mechanisms (Dennett, 2016).

Alongside philosophical exploration, neuroscientific research has made significant advances, particularly in identifying the neural correlates of consciousness (NCC) through sophisticated brain imaging technologies, such as fMRI and EEG (Dehaene et al., 2017). Prominent empirical frameworks such as Global Workspace Theory (GWT) and Integrated Information Theory (IIT) have garnered notable attention, proposing empirically testable hypotheses about the neural underpinnings of conscious experience. GWT, initially proposed by Bernard Baars and further refined by Stanislas Dehaene, suggests consciousness emerges when information becomes globally accessible to multiple cognitive processes through widespread neural networks, primarily involving frontoparietal circuits (Dehaene, Changeux & Naccache, 2011). Meanwhile, IIT, developed by Giulio Tononi, emphasises the intrinsic integration of information as central to consciousness, suggesting that conscious experiences correlate directly with a system's capacity to integrate information (Tononi et al., 2016). Recent empirical findings have pinpointed specific neural "hot zones," particularly in posterior cortical areas, that are significantly associated with conscious awareness, thereby enhancing our understanding of the neuronal basis of conscious states (Koch, Massimini, Boly, & Tononi, 2016).

The convergence between philosophical insights and neuroscientific discoveries has fostered an interdisciplinary dialogue, increasingly acknowledged as essential for meaningful progress. This synergy bridges the gap between conceptual rigour and empirical validation, informing experimental designs and refining theoretical propositions. Notably, interdisciplinary collaborations are instrumental in evaluating competing theories through rigorous empirical testing, exemplified by recent adversarial collaborations explicitly designed to differentiate between predictions from IIT and GWT (Mashour, Roelfsema, Changeux, & Dehaene, 2020).

The rapid advancement in artificial intelligence (AI) introduces an additional dimension to the discourse on consciousness, prompting critical inquiry into the criteria for machine consciousness. Recent theoretical analyses have systematically assessed whether current AI systems possess the requisite features for consciousness, concluding predominantly that present-day artificial systems lack essential elements, such as integrated information processing, global workspace mechanisms, and genuine experiential states (Chalmers, 2023; Butlin et al., 2023). Nonetheless, the possibility of future conscious AI remains a significant philosophical and ethical consideration, urging scholars to define more precise criteria for consciousness in artificial systems and explore the implications of potential machine consciousness.

This article critically surveys these multifaceted developments, synthesising philosophical debates, neuroscientific advancements, and considerations related to

Paper ID: SR251231121424 DOI: https://dx.doi.org/10.21275/SR251231121424

artificial intelligence. By systematically addressing contemporary controversies, methodological advances, and interdisciplinary integration, this work aims to enrich scholarly discourse and foster a cohesive, empirically grounded understanding of consciousness. In conclusion, continued collaborative research across philosophy, neuroscience, and artificial intelligence promises incremental advances and potentially transformative insights into one of humanity's most enigmatic phenomena.

2. Philosophical Theories of Consciousness

Qualia and the Hard Problem: At the heart of the mind-body problem is the existence of **qualia** – the subjective qualities of conscious experience (the redness of red, the pain of a headache). Philosophers agree that qualia *exist* in that we have subjective experiences, but they fiercely debate how to characterise them (Chalmers, The Conscious Mind: In Search of a Fundamental Theory., 1996). Chalmers famously dubbed explaining qualia "**the hard problem**" of consciousness – the challenge of why and how brain processes produce first-person experiences (Chalmers, The Conscious Mind: In Search of a Fundamental Theory., 1996, p. 4). Physicalist philosophers struggle to reduce or explain qualia in functional terms, while others suggest our current science may be incapable of bridging this explanatory gap. One response is **illusionism**, the view that qualia, as we intuitively conceive them, are not real properties. Illusionists argue that our brain generates *"a conjuring trick"* – it *seems* we have private qualitative sensations, but this is a cognitive illusion produced by complex neural processes (Frankish, 2016, pp. 11-12). In this view, there is "nothing it is like" to experience red; the brain simply makes us *believe* there is a non-physical quality (Frankish, 2016, pp. 14-15). Illusionism remains controversial, as many contend that denying the reality of phenomenal consciousness "begs the question" by dismissing what needs to be explained. The opposing stance holds that qualia are *real, irreducible* features of our mental life, which any complete theory of mind must account for (Chalmers, The Conscious Mind: In Search of a Fundamental Theory., 1996, pp. 95-96). This standoff underlies much of the modern philosophical debate on consciousness.

Panpsychism – Consciousness as Fundamental: In recent years, a significant trend in the philosophy of mind has been the resurgence of **panpsychism**. Panpsychism is the view that consciousness is a fundamental and ubiquitous property of matter, that even elementary particles or fields have proto-conscious aspects. (Goff, 2019, pp. 28-30). Rather than emerging *de novo* at some complex level of brain organisation, consciousness (in rudimentary form) is postulated to pervade the physical world. This idea, while radical, promises to "unitarily" resolve the mind-matter dualism by positing a single underlying substance with both physical and experiential aspects. (Goff, 2019, pp. 45-47). Advocates, such as philosopher Philip Goff (in *Galileo's Error*, 2019), argue that physics describes only the structural relations of matter from a third-person perspective. In contrast, consciousness describes the intrinsic nature of matter from the first-person perspective. Thus, every physical entity might have an "inside" aspect that feels like something, however primitive.

Panpsychism elegantly avoids the 'hard problem' by denying a sharp divide between matter and mind. There is no leap from non-conscious to conscious, since elementary awareness is built into everything. (Goff, 2019, pp. 46-48). However, critics highlight the **combination problem**: if atoms or neurons have micro-experiences, how do these combine to form the unified, complex consciousness that humans (or even mice) possess? Christof Koch observes, *"by what principle are [these] monadic boundaries decided?"* – e.g. why don't two people's minds ever merge into one, or how do billions of tiny conscious entities in a brain form a single coherent subjectivity (Koch C. , Consciousness: Confessions of a Romantic Reductionist., 2012, pp. 133-134)? Recent work has not fully solved this combination problem, and sceptics, such as neuroscientist Anil Seth, have criticised panpsychism for *"failing to explain how small conscious parts yield our consciousness"* and for lacking predictive power (Seth, 2021, pp. 53-55). Nonetheless, the panpsychism hypothesis has stimulated valuable discussion. It has even found surprising resonance with some neuroscientific theories (e.g., Integrated Information Theory), which treat consciousness as an intrinsic aspect of physical systems. (Tononi G. , 2015, pp. 11-13). Whether panpsychism is a profound insight or a "pseudo-explanation" is actively debated in philosophy journals (Searle, 2013, pp. 33-35), making it one of the most discussed theoretical frameworks of recent years.

Beyond Dualism – Contemporary Debates: Besides panpsychism, other philosophical perspectives have emerged in the quest to explain consciousness. **Dualist** interpretations (that the mind is non-physical) are now minority positions in academia, but the intuitions behind them fuel arguments that current materialist science might be missing something fundamental. On the other side, **reductive physicalists** hold that neuroscience will eventually explain consciousness *without* remainder – essentially solving the 'hard problem' by showing it was just an easy problem." Some propose refined physicalist accounts, such as **Russellian monism** (often allied with panpsychism), which suggests that matter has unknown intrinsic properties that could underlie consciousness, thereby reconciling dualism and materialism. Another active discussion is **higher-order theories** versus **first-order theories** of consciousness: higher-order theorists (like David Rosenthal) argue that a mental state is conscious only when one has thought about that state (a meta-representation), whereas first-order theorists believe direct representations can be conscious on their own. These debates are closely tied to empirical research on reflective awareness and self-consciousness (Rosenthal, 2005). Meanwhile, philosophers like Daniel Dennett and Keith Frankish champion *"illusionism"* (mentioned above) as the proper way to dissolve the 'hard problem', sparking ongoing exchanges about whether denying qualia solves or sidesteps the issue. In summary, recent philosophical discourse on consciousness has been vibrant, ranging from arguments about the existence and nature of qualia to proposals that consciousness is a fundamental property of the universe and to vigorous exchanges between those who view the 'hard problem' as a genuine mystery and those who consider it a conceptual illusion. This rich dialogue provides critical context and conceptual frameworks that guide (and sometimes challenge) the scientific investigations of consciousness.

3. Neuroscientific Advancements in Consciousness Research

Neural Correlates and Brain Imaging: On the empirical front, neuroscience has made significant strides in isolating the **neural correlates of consciousness (NCC)**—the specific brain processes that reliably correspond to conscious experience (Koch C. , 2004). Using tools such as functional MRI, EEG, and intra-cranial recordings, researchers compare brain activity when a stimulus is consciously perceived versus when it is not (for example, in visual masking experiments or inattentional blindness). **Recent studies** have identified several candidate signatures. For example, specific EEG waveforms, such as the P3b event-related potential and the *visual awareness negativity* (VAN), have been linked to conscious perception (Koivisto, 2010). High-frequency (gamma band) oscillatory activity and widespread cortical activation also tend to accompany conscious awareness of stimuli (Melloni L. M., 2007). Notably, a 2017 high-density EEG study by Siclari *et al.* demonstrated that when subjects reported dreaming (versus no experience) during sleep, there was a localised drop of low-frequency activity in a **"posterior hot zone"** of the cortex (Siclari, 2017). This posterior cortical region, specifically the temporal-parietal-occipital junction, has emerged as crucial for the contents of consciousness across states – it lights up during conscious perception and even during dream experiences, suggesting it may be a core neural correlate of consciousness (NCC) for subjective content (Koch C. M., 2016). Meanwhile, activity in the brainstem and thalamus (particularly the reticular activating system) is known to regulate the **level of consciousness** (wakefulness/arousal) but not specific conscious contents (Parvizi, 2001, pp. 139-143). This distinction between the **level** and **contents** of consciousness is now well-established: brainstem and subcortical circuits turn consciousness "on and off" (as in coma or general anaesthesia), whereas particular cortical networks underlie *what* one is conscious of at a given moment (Mashour, 2017, pp. 261-263). Over the last five years, advancements in neuroimaging and stimulation methods have enabled a more detailed examination of these mechanisms. For instance, intracortical electrical stimulation combined with EEG (PERT and PCI measures) can assess the brain's capacity for conscious integration—a crucial factor in evaluating covert consciousness in comatose patients. Overall, neuroscientific research has increasingly mapped the "footprints" of consciousness in the brain, finding that conscious experience correlates with a dynamic interplay of widespread cortical activation (ensuring information sharing) and specific localised activity patterns (defining content), all orchestrated in the context of an awake network able to sustain complex activity.

Global Workspace vs. Integrative Theories: A primary focus of recent empirical research is testing competing **theories of consciousness** that propose different neural mechanisms underlying consciousness. Two leading frameworks have dominated the discussion: **Global Workspace Theory (GWT)** and **Integrated Information Theory (IIT)**. GWT (formulated initially by Bernard Baars and refined by Stanislas Dehaene and others) posits that conscious awareness depends on a **global broadcasting** function in the brain (Baars, A Cognitive Theory of Consciousness. , 1988, pp. 15-17). In this view, many specialised processors in the brain operate unconsciously in parallel. However, when a particular piece of information "wins" attentional competition, it is loaded into a global workspace (often associated with frontoparietal circuits) where it becomes broadly available to other processes (memory, decision-making, language, etc.). In other words, mental content is conscious if and only if it is globally accessible – "broadcast" across the brain's network. (Baars, 2005, pp. 48-50). Neurally, this corresponds to a burst of synchronised activity (called neuronal ignition) involving the prefrontal and parietal cortices, enabling the information to impact many systems.

Evidence for GWT includes findings that tasks requiring reportable awareness consistently engage frontoparietal networks and that damage to these networks (or their disconnection) can impair conscious reports. However, GWT has been challenged by studies suggesting that, in some cases, the activity of primary sensory regions suffices for raw experience. This leads to a rival idea, often referred to as **Local Recurrent Processing** or **Recurrent Processing Theory (RPT)** (Lamme, 2006, pp. 494-496)**.** RPT (associated with Victor Lamme and others) argues that **re-entrant loops** of activity in sensory areas (e.g. visual cortex) can produce phenomenal consciousness even without frontal involvement – the frontoparietal ignition is only needed for cognitive access or report, not for the experience itself. This debate – whether frontal "global workspace" activity is necessary for the experience or only for its reporting – is an active area of research, with recent no-report paradigms (in which subjects' experiences are inferred without explicit reports) suggesting that some aspects of conscious perception occur even with minimal prefrontal activation. To adjudicate this, scientists have launched direct empirical **tests**. Notably, in 2021, an extensive adversarial collaboration (the *Cogitate* project) was initiated to pit GWT against IIT in carefully designed experiments. (Melloni L. M., 2021).

Integrated Information Theory (IIT): Unlike GWT's focus on access and broadcasting, **IIT (Integrated Information Theory)** presents a more fundamental and quantitative proposal. IIT (developed by Giulio Tononi and colleagues) suggests that what makes a system conscious is the degree to which it **integrates information**. It starts from phenomenological axioms (intrinsic existence, compositionality, information, integration, exclusion). It follows that a conscious experience corresponds to a single physical system that generates a particular amount of integrated information (denoted by the value **Φ**, "phi") (Tononi G. B., 2016, pp. 450-452). The higher the Φ of a system, the more unified and irreducible the system's internal causal structure, and thus the richer its conscious experience. According to IIT, a complex of neurons with high Φ (likely in the posterior cortex) constitutes the physical substrate of consciousness, whereas systems that are either too fragmented or too homogeneous (low Φ) will not have conscious experience (Koch C. M., 2016, pp. 453-455). Empirically, IIT-inspired research has identified the **posterior cortical hot zone** (encompassing parietal, temporal, and occipital regions) as generating high integration and being critical for the contents of consciousness (Koch C. M., 2016, pp. 308-310). This aligns with findings that disrupting the posterior cortex (with TMS or lesions) profoundly alters experience. In contrast, prefrontal lesions can sometimes leave core phenomenology intact (while affecting reporting or cognitive aspects). IIT has implemented measures such as the perturbational complexity index (PCI), which assesses consciousness in coma or vegetative patients by measuring brain integration. Recent studies comparing anaesthetised vs. awake states also support the idea that loss of consciousness correlates with a breakdown of integrated cortical dynamics (φ dropping). Nonetheless, IIT is controversial: it makes bold claims (even a simple photodiode might have a tiny consciousness if Φ>0, and conversely, a large AI lacking integration might be unconscious) that are debated. The adversarial collaboration explicitly tests predictions where GWT and IIT diverge. For example, IIT predicts consciousness can exist without global ignition in front areas and that certain stimuli will produce more posterior activation (integrated) even if not reportable. In contrast, GWT predicts that a lack of reports equals a lack of consciousness. Preliminary results of these tests have emerged: an initial report in 2023 suggested that some findings *slightly* favoured IIT's predictions over GWT's (Melloni L. M., 2023, pp. 3-4). However, the full verdict has not yet been reached. Importantly, these experiments reflect how neuroscience is now directly informed by philosophical theories, turning them into testable hypotheses. One paper noted that *"the neuroscience of consciousness is undergoing a significant empirical acceleration"* due to such theory-driven adversarial studies (Michel M. &., Minor disagreements and major disputes: The debates between the global neuronal workspace and integrated information theory of consciousness., 2021, pp. 2-3). Beyond IIT and GWT, many **other theories** exist (at least 22 supported neurobiological explanations exist, by one count (Michel M. , 2019, p. 3). For instance, **Higher-Order Thought (HOT)** theory posits that perception becomes conscious only when one has a higher-order representation of it, often linked to prefrontal cortex activity (Rosenthal, 2005, pp. 15-17). Another recent proposal, the **Attention Schema Theory (AST)** by Michael Graziano, suggests the brain's internal model of its attention processes gives rise to the subjective feeling of awareness. Moreover, in 2020, Andrew Budson and colleagues proposed a novel **"Memory Integration Theory"** of consciousness: the idea that what we experience as conscious perception is the brain's *memory system* binding and time-stamping incoming information, essentially a short-term memory of very recent events (Budson, 2020, pp. 4-6). This memory-centric view purports to explain odd temporal phenomena, such as postdictive effects (where later events influence how we perceive an earlier event) (Eagleman, 2000, pp. 389-391). Each theory attempts to explain experimental data and clinical observations, emphasising different aspects (attention, integration, higher-order representation, etc.). Distinguishing between these theories is now a key goal of the field. Efforts like the Theory Comparison project (TC) are developing standardised metrics to **quantify empirical support** for each theory, aiming to determine which frameworks best fit the growing body of data (Kirkeby-Hinrup A. &., 2023). In summary, neuroscience has progressed from merely identifying correlates of consciousness to rigorously testing **causal theories** of consciousness. The past five years have seen groundbreaking experiments, improved measures of brain complexity, and fruitful collaborations between scientists and philosophers to refine what such theories should explain. While a definitive theory is still elusive, this iterative feedback between theoretical proposals and empirical validation is steadily narrowing the field of viable explanations for how the brain generates conscious experience.

Bridging Philosophy and Neuroscience

Given the complexity of the problem, an **interdisciplinary approach** has become increasingly crucial in consciousness studies. Researchers recognise that purely philosophical or empirical approaches alone are insufficient; progress requires integrating conceptual clarity with scientific data. This has led to what is sometimes called **"neurophilosophy"** or simply "interdisciplinary consciousness science". As one recent paper noted, *"the field of interdisciplinary consciousness studies (ICS) has been blossoming… at the intersection of philosophy of mind, psychology, cognitive science, and neuroscience"* (Michel M. &., Minor disagreements and major disputes: The debates between the global neuronal workspace and integrated information theory of consciousness., 2021, p. 1). Dozens of competing theories have proliferated, and a purely empirical vote is not enough to decide between them – careful analysis of how evidence relates to each theory's claims is needed. Philosophers of science have stepped in to help design **frameworks for theory comparison**. For example, researchers are adopting formal approaches (inspired by Bayesian confirmation theory and Lakatos's philosophy of science) to evaluate how well the evidence supports various theories of consciousness. This ensures that when neuroscientists run experiments to test IIT vs GWT or other contenders, the interpretation of results is rigorous and theoretically informed.

Several high-profile initiatives underscore the bridging of disciplines. The **Templeton World Charity Foundation** recently funded large-scale adversarial collaboration experiments (as mentioned earlier) in which teams with opposing theoretical commitments jointly design studies to test their predictions fairly. (Michel M. (., 2022, pp. 3-4). Alongside neuroscientists, philosophers served as independent observers or judges, evaluating which theory's predictions matched the observed outcomes. Another example is the **Association for Mathematical Consciousness Science (AMCS)**, founded in 2021, which brings together experts in neuroscience, cognitive science, AI, and philosophy to develop mathematically precise theories of consciousness. In 2023, AMCS published an open letter calling for the responsible development of AI to *include* consciousness research, signed by notable figures from both AI (e.g., Yoshua Bengio) and neuroscience (e.g., Manuel Blum) (Community & signed by researchers including Yoshua Bengio, 2023). This letter exemplifies cross-disciplinary concern: advances in AI raise philosophical questions about mind and ethics, motivating scientific inquiry into the nature of consciousness.

Academic programs and conferences also foster this interdisciplinary dialogue. For instance, CIFAR's **Brain, Mind & Consciousness** program explicitly *"brings together neuroscientists, philosophers, and psychologists to grapple with the fundamental underpinnings of consciousness"*, linking biological findings to philosophical questions. (CIFAR, 2021). Likewise, the annual **Association for the**

Scientific Study of Consciousness (ASSC) meetings and the **Science of Consciousness** conferences (Tucson) feature philosophers and scientists side by side. Journals such as *Neuroscience of Consciousness* and *Philosophy and the Mind Sciences* publish work that bridges empirical results with theoretical analysis. (Michel M. &., 2021, p. 3). In recent special issues, authors have proposed criteria that any theory of consciousness must address, such as the *ontogenetic emergence* of consciousness during development. (Kirkeby-Hinrup A. &., 2023, pp. 2-4) Or its evolutionary function, drawing on both philosophical argument and neuroscientific evidence. Another fruitful bridge has been **neurophenomenology**, pioneered by Francisco Varela, which attempts to marry first-person phenomenological data with third-person neural data by having trained subjects provide fine-grained reports of experience to correlate with brain activity. While challenging, this approach aligns with the broader trend of treating subjective experience as a data source that can inform neuroscience rather than something to be ignored or explained away.

In summary, in the last five years, the barriers between philosophy and neuroscience of consciousness have continued to erode. Philosophy contributes vital clarity (e.g. framing what counts as an explanation, formulating the ‘Hard problem’, debating concepts like ‘illusion’ or ‘intrinsic nature’). In contrast, neuroscience provides objective constraints (e.g., no theory can ignore the empirical fact that the posterior cortex appears crucial or that specific brain injuries abolish consciousness). The mutual influence is evident: neuroscientists like Anil Seth engage with philosophy by proposing the “real problem” (explaining cognitive functions associated with consciousness), and philosophers like David Chalmers engage with neuroscience by speculating on testable indicators of consciousness. (Chalmers, 2020, pp. 5-6). Through interdisciplinary collaboration, the field aims to develop a **unified understanding** of philosophically coherent and empirically grounded consciousness. While a comprehensive solution remains distant, this bridge between disciplines is gradually paving the way toward a science of consciousness that respects the subjective richness of its subject matter while remaining firmly rooted in objective investigation.

Artificial Intelligence and Machine Consciousness

The rapid progress in artificial intelligence has sparked intense debate about **AI and consciousness**: Can machines ever be conscious, or can they simulate consciousness? What criteria would indicate machine consciousness? Furthermore, what are the ethical implications if they do (or even if they *almost* do)? Over the past five years, these questions have evolved from theoretical musings to urgent discussions, thanks partly to the development of advanced AI systems, such as deep neural networks and large language models. In 2022, for example, a Google engineer claimed that a conversational AI (LaMDA) had become sentient, provoking widespread controversy. While most experts were sceptical of that claim, it underscored the need for clear frameworks to assess machine consciousness.

Researchers have begun applying scientific theories of consciousness to AI systems to evaluate their potential for consciousness. A 2023 review by David Chalmers examined whether large language models (such as GPT-4) could be conscious by considering various indicators suggested by science. (Chalmers, 2023, pp. 6-10). These indicators include the ability to report on one’s internal states, the presence of recurrent (feedback) processing akin to human cortical loops, having some form of an embodied perspective or sensory grounding, possessing a global workspace architecture, and exhibiting unified agency or self-modelling (Chalmers, 2023, pp. 5-6). Chalmers concluded that current AI models *lack* virtually all these features – for instance, ChatGPT has no persistent self or sensory embodiment, no global broadcasting mechanism in the sense of GWT, and no genuine understanding of experiences – and therefore, it is **not plausible that they are conscious in their present form.** (Chalmers, 2023, pp. 25-27). Similarly, a comprehensive 2023 report by Butlin *et al.* surveyed leading theories (GWT, IIT, HOT, etc.). It derived a list of properties that each theory implies a system would need for consciousness. (Butlin, 2023, pp. 6-10). They then assessed current AIs against these properties. They concluded that **no existing AI system is a strong candidate for consciousness** – none of them meets the intersection of requirements across theories (for example, no AI today has the integrated causal structure with high Φ that IIT demands *or* the brain-like global workspace dynamics that GWT suggests) (Butlin, 2023, pp. 12-16). These scholarly analyses provide a more principled basis than earlier vague comparisons, and they largely agree that machine consciousness *either does not exist or, at the very least, has not been demonstrated*.

That said, researchers are actively exploring architectures that might one day produce machine consciousness or functional analogues. One notable effort is the work of **Manuel and Lenore Blum (2022)**, who proposed the “**Conscious Turing Machine (CTM)**,” a theoretical computer model influenced by Global Workspace Theory. The CTM is a simulated global workspace architecture featuring a memory, parallel processes, an attention mechanism, and a “spotlight” that broadcasts information to all processes, analogous to Baars' theatre of mind. The Blums demonstrated that such a system could mimic various cognitive phenomena associated with consciousness (e.g., it can model blindsight, where information is processed but not globally broadcast, resulting in no conscious report) (Blum, 2022, pp. 11-13). While the CTM does not *prove* the system is conscious (that remains a philosophical leap), it provides a concrete blueprint for how one might *engineer* a machine with functional properties of consciousness. Other AI researchers have drawn inspiration from neuroscience; for example, some have implemented simplified global workspace models in software agents (such as the ARCADIA system) to improve attention and self-reporting capabilities. (Cox, 2011, pp. 139-141). These implementations are steps toward AI that “knows what it is doing” in a human-like way, which some argue is a prerequisite for genuine consciousness and moral responsibility. (Shanahan, 2010, pp. 104-107).

A key question is whether **consciousness is substrate-independent** – can the exemplary cognitive architecture on a silicon computer yield consciousness, or is biological wetware exceptional? Most scientists are inclined to a functionalist view that substrate should not matter; only organisation does. However, IIT would suggest that a digital computer, which is almost always designed as a collection of discrete, feed-forward logic gates, might have very low Φ (because it is built to be modular and not irreducible), implying that standard computers *as currently built* might never achieve high consciousness even if they simulate it. This is speculative, but it illustrates the intersection of theory and engineering. Another aspect is **embodiment**: some argue that true consciousness requires a body and sensorimotor loops with the world (in line with philosophies of embodied cognition). Current language AIs lack physical embodiment or sensory modality, which may be why they lack a genuine self-model or first-person perspective. Efforts to integrate AI systems with robotic bodies or multimodal sensory inputs may become part of future attempts to create AI with more human-like awareness.

Besides the possibility, **the ethical debate** rages about desirability. If machine consciousness is possible, is it something we *want* to create? Some ethicists argue that a conscious AI would merit moral consideration (rights and protection from suffering), which complicates its use as a tool. Others, like philosopher Joanna Bryson, contend that we should *not* create conscious machines, famously stating, *"Robots should be slaves"*, and thus should not be endowed with capacities that make them suffer or grant them moral rights. (Bryson, 2010, pp. 63-65). There is also an argument that specific advanced AI could become moral agents *only if* they are conscious (because only then can they genuinely understand and intend actions) (Gunkel, 2012, pp. 129-132). This has implications for AI safety: some suggest that a non-conscious AI might be easier to align since it would not have genuine desires or a survival instinct, whereas a conscious AI might develop self-preservation. In 2023, the discussion reached policy circles, with the aforementioned AMCS open letter urging that **consciousness research be included in AI development agendas** to anticipate these issues. (AMCS, 2023).

In summary, the intersection of AI and consciousness research has evolved rapidly over the last five years. On the one hand, **most experts do not believe that current AI systems are conscious**, and systematic analysis using established theories of consciousness supports this view (Butlin et al., 2023; Chalmers, 2023). On the other hand, the **path toward possible machine consciousness** is being sketched out through theoretical models (such as the conscious Turing machine) and brain-inspired architectures. This research is inherently interdisciplinary: it involves computer scientists, neuroscientists, and philosophers collaborating to define what it would entail for a machine to be conscious and how we would recognise it. As AI capabilities continue to evolve, this debate is likely to intensify. In the coming years, prototype AI agents with rudimentary self-models or global workspaces may be used experimentally to probe whether such systems exhibit any signs of minimal consciousness. Even if they do not, building them will enrich our understanding of consciousness by testing the sufficiency of various mechanisms. Thus, the dialogue between AI engineering and consciousness science could be mutually illuminating: using consciousness theories to inform AI design and using AI models as test beds for consciousness theories. For now, machine consciousness remains a theoretical possibility that we are only beginning to systematically explore using the hard-won insights from both philosophy and neuroscience of consciousness.

4. Conclusion

Over the last five years, consciousness research has advanced on multiple fronts, yielding a more nuanced- albeit still incomplete- understanding of this profound phenomenon. Philosophers have sharpened the theoretical landscape, debating whether consciousness is an irreducible part of nature or an emergent property that cognitive functions can explain. Notions like qualia and the 'hard problem' remain central, with ongoing debates between those who view qualia as the undeniable essence of the mind and those who would explain them away as illusions. At the same time, novel philosophical proposals (e.g., panpsychism) have challenged orthodoxy by suggesting that consciousness pervades the universe at every level, forcing scientists to consider assumptions beyond the standard physicalist framework. Neuroscience, for its part, has delivered an ever-growing catalogue of empirical findings: specific brain signatures of conscious states, identified "hot zones" of cortical activity linked to experience, and refined tools to measure the brain's integrative capacity for consciousness Large-scale collaborations are actively testing prominent theories, such as Global Workspace and Integrated Information Theory, against each other in rigorous experiments, an unprecedented empirical venture in a field that once struggled to find traction in the lab. The results of these tests, along with new theoretical insights, are gradually shaping a consensus on which aspects of brain activity are essential for consciousness (e.g., some degree of global informational integration seems critical, whether achieved via frontoparietal broadcasting, posterior connectivity, or both). Significantly, the divide between philosophical and scientific approaches to consciousness has narrowed. Interdisciplinary frameworks ensure that experimental design and theory evaluation proceed hand-in-hand, leveraging philosophical rigour to interpret empirical data and vice versa. This synergy is evident in how consciousness science now tackles questions once thought purely philosophical (such as the criteria for consciousness or the possibility of it in non-biological systems) with empirical seriousness.

In artificial intelligence, what was formerly science fiction has become a serious topic of scholarly inquiry. Experts have begun outlining the requirements for **machine consciousness** by drawing on established theories. Although current AIs do not meet these requirements, defining them has been valuable for clarifying our understanding of consciousness. It has underscored, for example, the importance of features like embodiment, self-monitoring, and integrated cognitive architectures, which are also themes in human consciousness research. The debate over AI consciousness has also had a reciprocal effect, prompting researchers to ask, "What *exactly* about the human brain makes us conscious, and can it be abstracted?" Answering this may inform AI design and illuminate why specific brain structures (and perhaps not others) give rise to experience.

In conclusion, today's study of consciousness is a dynamic convergence of philosophy, neuroscience, psychology, and computer science. **Recent philosophical work** has refined the questions and proposed bold new answers (e.g., treating consciousness as fundamental). Empirical **neuroscience** has delivered rich data and increasingly powerful tests of theories, and **bridging efforts** have created a dialogue that ensures these domains inform each other. While a comprehensive explanatory theory of consciousness remains out of reach, progress over the past five years is noteworthy: We now have clearer theoretical options, better experimental paradigms, and a more collaborative and interdisciplinary community than ever before. The *mystery* of consciousness is steadily chipped from both ends – conceptually and empirically. As this trend continues, we can be cautiously optimistic that each new study or debate, whether it narrows down the neural substrates or sharpens a philosophical argument, brings us closer to understanding how the miracle of conscious awareness arises from the workings of the natural world.

References

[1] Association for Mathematical Consciousness Science (2023). Open Letter: The responsible development of the AI agenda must *include consciousness research*. AMCS Website. https://amcs-community.org/open-letter/

[2] Bryson, J. J. (2010). *Robots should be slaves. In Y. Wilks (Ed.),* Close Engagements with Artificial Companions. John Benjamins Publishing, pp. 63–74

[3] Butlin, P., Shevlin, H., Mogensen, A., & Williams, D. (2023). *Consciousness in Artificial Intelligence: Insights from the Science of Consciousness.* UK Government Office for Science. https://www.gov.uk/government/publications/consciousness-in-artificial-intelligence

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

[5] Chalmers, D. J. (2020*). The meta-problem of consciousness*. Journal of Consciousness Studies. 27(5–6), pp. 5–35

[6] Chalmers, D. J. (2023). *Could a large language model be conscious*? Journal of Consciousness Studies. 30(5–6), pp. 6–33

[7] CIFAR (2021). *Brain, Mind & Consciousness Program*. CIFAR. https://cifar.ca/research/programs/brain-mind-consciousness/

[8] Cox, M. T., & Rajapakse, D. C. (2011). *A model of planning for intelligent agents based on the cognitive architecture ARCADIA*. Cognitive Systems Research. 12(3–4), pp. 137–157

[9] Gunkel, D. J. (2012). *The Machine Question: Critical Perspectives on AI, Robots, and Ethics.* MIT Press. Cambridge, MA

[10] Kirkeby-Hinrup, A., & Overgaard, M. (2023). *Ontogenetic emergence as a criterion for theories of consciousness.* Philosophy and the Mind Sciences. 4, Article 9902, pp. 1–20

[11] Lau, H., & Michel, M. (2020*). A framework for evaluating consciousness theories*. Philosophy and the Mind Sciences. 1(II), Article 1, pp. 1–24

[12] Michel, M., & Morales, J. (2021). *Minor disagreements and major disputes.* Neuroscience of Consciousness. 2021(1), niab001

[13] Seth, A. K. (2016). *The real problem.* Aeon Essays. https://aeon.co/essays/the-hard-problem-of-consciousness-is-a-distraction-from-the-real-one

[14] Shanahan, M. (2010). *Embodiment and the Inner Life.* Oxford University Press. Oxford

[15] Siclari, F. et al. (2017). *The neural correlates of dreaming.* Nature Neuroscience. 20(6), pp. 872–878