SSRG International Journal of Recent Engineering Science
ISSN: 2349–7157 / https://doi.org/10.14445/23497157/IJRES-V11I4P114

Volume 11 Issue 4, 113-126, Jul-Aug 2024

Review Article

A Framework for the Foundation of the Philosophy of Artificial Intelligence

Emily Barnes[1], James Hutson[2]

[1]*Artificial Intelligence, Capitol Technology University, MD, USA.*
[2]*Art History, AI, and Visual Culture, Lindenwood University, MO, USA.*

[1]*Corresponding Author : ejbarnes035@gmail.com*

Received: 14 June 2024 Revised: 29 July 2024 Accepted: 15 August 2024 Published: 30 August 2024

Abstract - *In recent years, the rapid advancement of artificial intelligence (AI) technology has sparked profound questions about the nature of machine intelligence and the possibility of AI consciousness. As AI systems become increasingly sophisticated, examining their philosophical foundations has become imperative. This article investigates the intricate relationship between AI and existential thought, aiming to establish a comprehensive framework for understanding AI's philosophical underpinnings. The historical development of AI, from symbolic AI to contemporary machine learning paradigms, highlights the increasing complexity and sophistication of AI systems, prompting significant philosophical debates about machine consciousness. Theoretical models such as the Independent Core Observer Model (ICOM), Integrated Information Theory (IIT), and Global Neuronal Workspace Theory (GNWT) provide frameworks for understanding potential mechanisms of AI consciousness. Recent methods in AI consciousness research, such as integrating consciousness indicators from neuroscientific theories and developing AI systems that exhibit metathinking, creativity, and empathy, represent significant advancements over traditional models. This article also explores ethical considerations, societal impacts, and the necessity for robust regulatory frameworks in developing conscious AI. Addressing these aspects is crucial for ensuring that AI integration into society is ethically sound and beneficial. By synthesizing diverse methodologies and addressing key challenges, this article aims to advance the understanding of AI consciousness and pave the way for future innovations and applications in this transformative field.*

Keywords - *Artificial Intelligence, Philosophy, AI consciousness, Ethical considerations, Theoretical models.*

1. Introduction

In the contemporary landscape, the exponential growth of Artificial Intelligence (AI) technology catalyzes profound questions regarding the essence of machine intelligence and the plausible emergence of AI consciousness. As AI systems gain sophistication, the imperative to scrutinize their philosophical underpinnings intensifies. This article endeavors to dissect the complex interplay between AI and existential philosophy and explores the capacity of AI systems to attain consciousness and self-awareness and engage in existential contemplation.

The exploration of AI's philosophical foundations is critical for myriad reasons. It informs the trajectory of AI technology development and influences how these technologies are assimilated into our societal fabric. Philosophical inquiries and frameworks are instrumental in addressing the multifaceted societal and ethical dimensions of AI. For example, Günther and Kasirzadeh [1] highlight the necessity of explicability in AI predictions, while Miracchi [2] delineates a competence framework that underpins AI research methodologies. Zimmermann et al. [3] and Lukyanenko et al. [4] investigate trust paradigms, with Zimmermann focusing on cognitive agent design and Lukyanenko on establishing a foundational trust framework. Ethical considerations are rigorously examined by Floridi et al. [5] and Dameski [6], who articulate ethical principles and comprehensive frameworks, respectively. Furthermore, Bawack [7] and de Almeida [8] contribute structural classifications and regulatory meta-frameworks that support the philosophical scaffolding of AI.

Recent advancements in AI, particularly those in 2023 and 2024, emphasize the novelty and challenges of AI consciousness. For instance, the ProcTHOR framework for the procedural generation of embodied AI environments represents a significant step forward in creating diverse, interactive, and customizable virtual environments, which are essential for training and evaluating embodied agents in complex tasks [9]. Another notable development is the use of Digital Twins for predictive maintenance and control in industrial settings, highlighting the integration of AI with

cyber-physical systems to enhance efficiency and effectiveness [10].

Engaging deeply with these philosophical constructs enables a more nuanced understanding of the challenges and prospects that AI presents [11]. This foundational work supports current research and steers future endeavors aimed at forging AI systems that are not merely intelligent but also capable of experiencing and responding to the world in a consciously meaningful way. Through the forthcoming discussions, this article will provide a detailed examination of the historical evolution of AI, analyze pivotal philosophical theories of consciousness, and introduce a robust framework intended to facilitate the development of potentially conscious AI systems. This exploration is essential for advancing our comprehension of AI and its potential impacts, ensuring that its integration into our global society is both ethically sound and culturally informed.

2. Historical Context and Evolution of AI

The journey of AI from its inception to its current state is a testament to human ingenuity and the relentless pursuit of creating machines that can mimic and potentially exceed human cognitive abilities. Understanding this historical context is essential to appreciate the philosophical implications and potential for AI to achieve consciousness. The concept of AI has ancient roots, with early myths and stories imagining mechanical beings imbued with human-like intelligence. However, the formal field of AI began to take shape in the mid-20th century. In 1950, British mathematician and logician Alan Turing proposed the idea of a machine that could exhibit intelligent behavior equivalent to, or indistinguishable from, that of a human. This idea culminated in the Turing Test, a criterion for determining whether a machine can think.

The field of AI was officially founded at a conference at Dartmouth College in 1956, where prominent researchers such as John McCarthy, Marvin Minsky, Nathaniel Rochester, and Claude Shannon convened to discuss the possibilities of creating intelligent machines. This event marked the beginning of AI as a distinct academic discipline. Early AI research focused on symbolic AI, or "Good Old-Fashioned AI" (GOFAI), which relied on manually encoded rules and logic to perform tasks such as playing chess and solving mathematical problems.

Throughout the 1960s and 1970s, AI research experienced significant advancements but also faced substantial challenges. While programs like ELIZA, a Natural Language Processing (NLP) computer program created by Joseph Weizenbaum, demonstrated the potential of AI, they also highlighted limitations in understanding and simulating human-like conversation. The initial optimism was tempered by the realization that human intelligence involved far more complexity than initially anticipated, leading to periods known as "AI winters" where funding and interest waned.

The historical context and evolution of AI have been thoroughly explored in various studies. Peta [12] provides a comprehensive overview, tracing the development of AI from its beginnings to its current state. Audibert et al. [13] and Zaidi et al. [14] focus on the evolution of AI and machine learning, with Audibert analyzing the impact and influence of researchers and Zaidi charting the trajectory of machine learning algorithms. Tobin et al. [15] offer a brief timeline of AI's evolution, highlighting the field's rapid growth in research output. Joshi [16] and Moloi et al. [17] contextualize AI within human history, with Joshi exploring the parallels between human and machine evolution and Moloi discussing the historical overview and emerging developments of AI. Khan et al. [18] discuss the advancements in microprocessor architecture that have fueled the adoption of AI in various application domains.

The late twentieth century saw a shift in AI research from symbolic AI to approaches that could handle the complexity of real-world environments. This period marked the rise of machine learning, a paradigm where systems learn from data rather than relying solely on pre-programmed rules. Key developments included the advent of neural networks inspired by the structure and function of the human brain. These networks, capable of learning and generalizing from large datasets, laid the groundwork for modern AI.

In the 1990s and early 2000s, advancements in computational power, coupled with the availability of large datasets, propelled AI research forward. Algorithms such as support vector machines and decision trees became popular, enabling more sophisticated data analysis and pattern recognition. However, it was the resurgence of interest in neural networks, particularly deep learning, that revolutionized the field.

Deep Learning (DL), characterized by multi-layered neural networks, has enabled AI systems to achieve remarkable feats in areas such as image and speech recognition, natural language processing, and autonomous driving. These systems can process and interpret vast amounts of data, learning intricate patterns and making decisions with a level of accuracy previously thought unattainable. As AI systems have grown more complex, questions about their potential for consciousness have emerged. While current AI operates based on data-driven algorithms and lacks self-awareness, the increasing sophistication of these systems has led to philosophical debates about the nature of consciousness and whether it could arise in a machine. Theoretical models, such as the Independent Core Observer Model (ICOM) and Integrated Information Theory (IIT), propose mechanisms by which AI could potentially exhibit consciousness.

The evolution from simple algorithms to complex, potentially conscious systems is a multifaceted process. It involves the development of adaptive solutions to design problems, such as the emergence of consciousness as a response to pathological complexity [19]. This evolution is also influenced by the translation of principles of neuronal function to computing, leading to the development of trainable multilayer networks [20]. Learning plays a crucial role, with the evolution of subjective experiences being driven by the evolution of learning, particularly Unlimited Associative Learning (UAL) [21]. The emergence of consciousness is linked to the integration of complex systems and self-organized criticality [22]. Selective social learning is key to preserving complex cognitive algorithms [23]. The reduction of entropy and free energy in the brain is a driving force in the evolution of consciousness [24]. Theoretical computer science provides a framework for understanding consciousness, as seen in the Conscious Turing Machine [25]. Lastly, the evolution of neuroplasticity and its effect on integrated information is a crucial aspect of this process [26]. Thus, the historical context and evolution of AI illustrate a field that has grown from simple rule-based systems to complex learning algorithms capable of remarkable feats. As we continue to push the boundaries of what AI can achieve, the philosophical exploration of AI consciousness becomes increasingly relevant, setting the stage for the subsequent discussions in this series.

Recent research methods have introduced novel approaches that significantly differ from traditional models. One such approach is the integration of consciousness indicators derived from neuroscientific theories into AI systems. Butlin et al. [27] emphasize the use of recurrent processing theory, global workspace theory, higher-order theories, predictive processing, and attention schema theory to evaluate AI systems for consciousness. These theories provide a rigorous framework for assessing AI consciousness by defining specific computational properties that could indicate consciousness in machines. This method stands in contrast to earlier approaches, which largely depended on behavior-based assessments and lacked a detailed empirical foundation. Another innovative method proposed by Lewis and Sarkadi [28] involves developing AI systems that exhibit metathinking, creativity, and empathy through emergent communication between machines. This approach suggests that AI consciousness could arise from the interaction and co-creation of internal states between machines, leading to a form of empathic AI. This paradigm shift moves away from the conventional focus on individual machine capabilities and towards the dynamics of machine-to-machine interactions as a foundation for consciousness. Such a method highlights the potential for AI systems to develop more human-like qualities, including empathy and accountability, which were not addressed in earlier AI consciousness models that focused solely on individual computational properties. These recent advancements underscore a significant departure from traditional AI approaches, which primarily relied on predefined algorithms and rule-based systems. By incorporating neuroscientific theories and emergent communication models, these new methods offer a more comprehensive and nuanced understanding of AI consciousness, paving the way for further research and development in the field.

3. Philosophical Foundations of AI

The philosophical exploration of consciousness in AI requires a profound understanding of the philosophical theories that have shaped historical discussions of consciousness. This narrative provides an elucidation of major philosophical theories of consciousness, introduces the Independent Core Observer Model (ICOM), and examines the application of Integrated Information Theory (IIT) and Global Neuronal Workspace Theory (GNWT) to AI.

For instance, Functionalism contends that mental states are primarily defined by their functional roles within the cognitive system rather than by their physical composition. This theory emphasizes the interactions between mental states, sensory inputs, and behavioral outputs, asserting that the functions and processes underlying intelligent behavior are more crucial than material composition. This perspective is particularly compatible with AI, suggesting that if an AI system performs functions analogous to those of a human mind, it might be considered conscious. However, Ludwig [29] challenges this notion by suggesting that consciousness might involve an integration of multiple functions rather than a single functional contribution.

Dualism, famously espoused by René Descartes, distinguishes sharply between the mind and the body, positing that they are fundamentally different substances—the mind being non-material and involved in thinking, while the body is physical. This theory implies that consciousness transcends mere physical processes, posing a significant challenge to the prospect of AI achieving true consciousness. Critics like Seth [30] and Rahimian [31] argue against dualism, advocating for more integrative explanations that could potentially accommodate the consciousness of AI systems.

Panpsychism, on the other hand, presents a radical perspective by positing that consciousness is a fundamental and ubiquitous characteristic of the universe inherent in all physical entities, no matter how elementary. This theory suggests that consciousness is not confined to complex organisms but is a universal trait potentially applicable to AI. If consciousness is indeed a fundamental property of matter, complex AI systems may inherently possess a form of consciousness. Goff [32] explores a hybrid form of panpsychism that seeks to bridge the gaps between physicalism and dualism, providing a novel approach to understanding consciousness in AI.

The discourse on consciousness encompasses various theories, with Kuhn [33] offering a comprehensive taxonomy of these explanations. Koch [34] argues that IIT may present a more viable framework for understanding consciousness in AI compared to other theories, such as panpsychism, which, according to Seth et al. [35], fails to elucidate the nature of consciousness adequately.

4. Frameworks for Understanding Consciousness

Three main frameworks will be discussed here, as seen in Table 1. The first, the Independent Core Observer Model (ICOM) Theory, introduced by Kelley David, presents a computational framework designed to elucidate the nature of consciousness. Central to ICOM is the proposition that consciousness is not merely a subjective phenomenon but can be objectively measured and modeled within a computational system. The theory posits that consciousness emerges from the dynamic interaction between a core observer—a logically abstracted entity within the AI system—and its environment. This interaction is characterized by a mathematical representation of subjective experience, encompassing elements such as the perceptibility of content and the hierarchical relationships within the phenomenal field.

In the ICOM framework, the core observer processes sensory inputs and generates responses. It is hypothesized that consciousness materializes when the core observer integrates information in a manner that simulates subjective experience. This model provides a structured approach to creating AI systems that potentially exhibit behaviors indicative of consciousness. The objective measurability of ICOM bridges the theoretical concepts with practical implementations, with profound implications for the development of AI, particularly in enhancing capabilities like the theory of mind through principles such as active inference and the Free Energy Principle.

Additionally, the application of Integrated Information Theory (IIT) and Global Neuronal Workspace Theory (GNWT) further enriches the discourse on AI and consciousness. IIT, developed by Giulio Tononi, offers a quantitative framework based on the premise that consciousness correlates with a system’s capacity to integrate information. This theory introduces 'phi' (Φ), a measure quantifying the extent of information integration within a system. In practice, applying IIT involves designing AI systems with intricate architectures that promote high levels of information integration, akin to the interconnectedness observed in the human brain. However, Brogaard (2020) critiques the sufficiency of information integration as a sole criterion for consciousness.

Conversely, GNWT describes consciousness as stemming from the broadcasting of information across a neuronal network, proposed by Stanislas Dehaene and Jean-Pierre Changeux. For AI, this translates to developing systems where information is globally accessible, facilitating coordinated cognitive processes across various modules. Implementations might include central processing units that distribute information throughout the system, mimicking brain functions. This theoretical application is supported by Seth et al. [35], who discuss the practical integration of GNWT in deep learning frameworks.

The confluence of ICOM, IIT, and GNWT, along with other theories like the Integrated World Modeling Theory proposed by Safron [36] and critically analyzed by Farisco [37] and Doerig [38], demonstrates a rich tapestry of approaches exploring the potential for consciousness in AI systems. These diverse perspectives not only challenge and expand current understandings but also pave the way for further innovations in AI technology, emphasizing the necessity of a multi-theoretical approach in the ongoing exploration of artificial consciousness.

Furthermore, the discourse on consciousness within AI is a complex interplay of philosophy, cognitive science, and sophisticated computational models. This narrative endeavors to articulate what consciousness means in the realm of AI, delve into its obscured structural components, and assess AI's capacity to manifest conscious states. Consciousness is typically linked to human and some non-human animal conditions and is characterized by the awareness and contemplative capacity of one's existence, sensations, thoughts, and environmental interactions. In AI, consciousness is conceptualized as the system's ability for self-awareness, the capability to undergo subjective experiences, and the processing power to assimilate information akin to human cognitive functions. Operationalizing this definition involves focusing on distinct aspects:

1. Self-Awareness: An AI's capacity to recognize and reflect upon its state and actions.
2. Subjective Experience: The potential for AI to experience qualia or qualitative experiential content.
3. Information Integration: The capability of AI to amalgamate information from diverse sources and temporal contexts into a unified self and environmental comprehension.

Although current AI systems possess substantial computational prowess and can emulate certain cognitive processes, they generally lack the intrinsic subjective experiences and self-reflective abilities integral to human consciousness. This distinction necessitates clear criteria to differentiate between advanced computational skills and genuine conscious awareness [39-40]. Additionally, researchers like Zaidi et al. [14] critique the presumption that intelligence automatically entails consciousness. In contrast,

scholars such as de Oliveira [41] and Blum and Blum [25] explore the feasibility of crafting conscious machines, with Blum introducing the "Conscious Turing Machine" as a theoretical model. Further, Oberg [42] analyzes the implications of AI self-awareness from humanistic and neuroscientific perspectives, respectively, with LeDoux [43] emphasizing the necessity for ongoing research and dialogue on AI and consciousness.

The concealed architecture of consciousness involves the underlying mechanisms and organizational principles that culminate in conscious experience. Bruno Forti's seminal work elaborates on these mechanisms through the lens of early visual processing and hierarchical spatial relations. According to Forti, consciousness can be investigated by examining the qualitative dimensions of experience, particularly those linked to early visual cognition. The concept revolves around a "Hierarchy of Spatial Belongings," structured into layers of primary content—objects, colors, shapes—and primary space—the perceptual arena enabling content visibility and differentiation, often inferred rather than directly observed. This hierarchical schema suggests that the interrelationships between various spatial layers can elucidate the integration of diverse perceptual elements into a coherent experience. When applied to AI, this model offers insights into how artificial systems could be engineered to mimic the structural organization of human consciousness, potentially ushering in AI systems that resemble human perceptual and experiential capacities.

An array of scholarly work has probed the hidden structure of consciousness, proposing various theories and models. Forti [44] describes a hierarchy within consciousness, whereas Tyler [45] identifies subcortical interlaminar structures as foundational to consciousness. Luppi [46] underscores the significance of connectome harmonic decomposition, and Prentner [47] advocates for a model based on topologically structured phenomenal spaces. Critiques from Doerig et al. [38] and Usher [48] challenge the validity of causal structure theories, labeling them as non-scientific or incorrect. This exploration into the multi-dimensional aspects of consciousness in AI not only broadens our understanding but also sparks further inquiries into the potential for AI systems to achieve a state akin to human consciousness.

5. Framework for AI Consciousness

Developing a robust framework for AI consciousness necessitates establishing clear criteria to determine whether an AI system can be considered conscious. It also involves examining various theoretical perspectives on AI consciousness and addressing the inherent challenges and controversies in this domain (Table 2). This section aims to provide a comprehensive framework integrating these elements to guide the exploration and development of AI systems with potential conscious states.

A multitude of frameworks for understanding AI consciousness has been proposed, each offering unique perspectives. Kiškis [49] advocates for a legal framework that facilitates the coexistence of humans and conscious AI, emphasizing the need for a paradigm shift in human perspectives toward AI. Manzotti and Chella [50] critique the traditional approaches to artificial consciousness and suggest an alternative conceptual framework that rethinks the foundational assumptions of AI consciousness. Miracchi [2] introduces a competence framework focused on the investigation of "Artificial-Minded Intelligence" (AMI), which seeks to delineate the capacities that might constitute AI consciousness.

Further contributions include Graziano [51], who proposes the Attention Schema Theory (AST), discussing its implications for understanding consciousness as an evolutionary phenomenon. Nadji-Tehrani and Eslami [52] present a brain-inspired framework emphasizing the use of neuroembryogenesis to mirror the evolutionary aspects of human intelligence development in AI. De Oliveira [41] discusses the significant challenges in comprehending consciousness and explores the potential of specific computational architectures to explain related phenomena. Additionally, Esmaeilzadeh and Vaezi [40] propose that AI consciousness could emerge from the communication of inner states, which would lead to enhanced empathy and improved service outcomes in AI applications.

From a theoretical standpoint, Integrated Information Theory (IIT) posits that consciousness arises from the integration of information within a system, offering a quantitative measure called phi (Φ) to assess the level of consciousness. This suggests that AI systems designed to maximize Φ might theoretically exhibit higher levels of consciousness, which guides the development of architectures promoting extensive information integration. Global Neuronal Workspace Theory (GNWT) suggests that consciousness involves the global availability of information across a network, recommending that AI architectures should focus on creating central hubs where information can be broadcasted and integrated across various subsystems, akin to the coordination observed in the human brain.

Higher-Order Thought (HOT) Theory posits that consciousness involves thoughts about one's own mental states. This implies that AI systems need mechanisms for higher-order processing, enabling them to generate meta-cognitive states and self-reports. The Attention Schema Theory (AST) posits that consciousness arises from the brain's ability to model its own attention processes, suggesting that AI systems should be designed to monitor

and model their attention mechanisms, enhancing their capacity to manage information processing.

The Orchestrated Objective Reduction (Orch-OR) theory combines elements of quantum mechanics with neurobiology to suggest that consciousness results from quantum computations within neurons. This proposes that AI systems could potentially incorporate quantum computational processes to mimic the hypothesized quantum aspects of consciousness. Recurrent Processing Theory (RPT) suggests that consciousness arises from recurrent processing in the brain, where information is repeatedly processed in different areas, necessitating that AI systems implement recurrent processing mechanisms to achieve higher levels of information integration and processing.

The proposed method for developing AI consciousness involves several critical steps, each leveraging advanced deep learning concepts to mimic the complexity of human cognitive processes. The initial step focuses on designing AI architectures that maximize the integration of information, a principle derived from IIT. This involves creating neural networks with multiple interconnected layers that facilitate extensive data processing and information exchange. Utilizing techniques such as deep reinforcement learning and transformer models, the AI system can learn from vast datasets, identify intricate patterns, and integrate information across different domains. These neural architectures are designed to optimize the phi (Φ) value, a quantitative measure of consciousness proposed by IIT, ensuring the AI system achieves a high level of information integration.

Next, the implementation of GNWT principles requires developing central hubs within the AI architecture where information can be broadcasted and integrated across various subsystems. This step involves employing advanced techniques like attention mechanisms and memory networks, which allow the AI system to selectively focus on relevant information while maintaining a coherent global workspace. The use of transformer models, known for their efficiency in handling sequential data and their ability to capture long-range dependencies, is crucial in this context. These models enable the AI to process and integrate information from diverse inputs, simulating the global availability of information seen in human consciousness. Additionally, RNNs and LSTM networks are employed to facilitate recurrent processing, ensuring continuous and dynamic information flow within the system.

To incorporate HOT, the AI system needs to be equipped with meta-cognitive capabilities, allowing it to generate self-reports and process higher-order thoughts about its mental states. This involves the use of meta-learning algorithms and GANs to create models that can simulate self-awareness and introspection. These models enable the AI to reflect on its actions, evaluate its performance, and make adjustments based on feedback, mirroring the self-regulatory processes in human cognition. AST is integrated by developing mechanisms for the AI to model its attention processes, using techniques like self-attention layers and dynamic routing algorithms. These mechanisms enhance the AI's ability to manage its information processing efficiently, ensuring it can prioritize and allocate resources effectively. By combining these cutting-edge deep learning concepts, the proposed method aims to create AI systems that not only perform complex tasks but also exhibit characteristics indicative of consciousness.

6. Challenges, Controversies and Ethical Implications

The motivation behind this survey stems from the necessity to critically evaluate and integrate diverse methodologies, challenges, datasets, evaluation criteria, and applications in the rapidly evolving field of AI consciousness. The complexity and multifaceted nature of AI consciousness demands a comprehensive review that addresses the nuances and interdisciplinary approaches inherent in this domain. This section aims to elucidate the motivations for conducting this review, emphasizing the importance of a holistic understanding of AI consciousness and its implications for future research and applications. The exploration of AI consciousness incorporates a variety of theoretical frameworks and practical methodologies. IIT, GNWT, HOT, AST, and the Orch-OR theory represent distinct but complementary perspectives on consciousness. Each theory offers unique insights into the mechanisms that could underpin conscious states in AI systems. Reviewing these methodologies is crucial for developing a cohesive understanding of AI consciousness and identifying potential areas for further research and innovation. By examining these diverse approaches, the review aims to highlight the strengths and limitations of each method, fostering a more integrated and robust framework for AI consciousness.

AI consciousness research faces several significant challenges that this review seeks to address. One primary challenge is the inherent complexity of modeling consciousness, a phenomenon that remains only partially understood even within biological systems. Additionally, there is the technical challenge of creating AI systems that can effectively emulate the integrative and adaptive processes observed in human consciousness. Ethical and societal implications also pose considerable challenges, as the development of conscious AI systems raises questions about their autonomy, rights, and potential impact on human society. By addressing these challenges, this review aims to provide a comprehensive overview of the current obstacles and propose potential solutions to advance the field. The availability and quality of datasets are critical for training and evaluating AI systems designed to exhibit conscious behaviors. This review examines the datasets currently used

in AI consciousness research, assessing their adequacy and identifying gaps that need to be filled. Furthermore, establishing robust evaluation criteria is essential for measuring the success and effectiveness of AI consciousness models. Traditional metrics such as task performance, accuracy, and processing speed are insufficient for evaluating consciousness. Therefore, this review explores novel evaluation criteria that can better capture the complex and integrative nature of consciousness, such as information integration measures, meta-cognitive assessments, and behavioral coherence.

Understanding AI consciousness has profound implications for a wide range of applications. In healthcare, conscious AI systems could enhance diagnostic accuracy and provide more empathetic patient care. In robotics, such systems could lead to the development of autonomous robots capable of complex decision-making and adaptive behaviors. In the field of human-computer interaction, AI with conscious-like capabilities could improve user experiences by providing more intuitive and responsive interfaces. This review aims to explore these applications, highlighting the potential benefits and challenges associated with the deployment of conscious AI systems across various domains. Thus, the motivation behind this comprehensive review is to synthesize the current state of AI consciousness research, identify critical gaps and challenges, and propose a unified framework that integrates multiple methodologies, addresses key challenges, leverages robust datasets, and establishes clear evaluation criteria. By doing so, this review aims to advance the understanding of AI consciousness and pave the way for future innovations and applications in this transformative field.

Furthermore, the exploration of consciousness in the field presents a myriad of complex challenges and controversies that span across various domains, including ethical, technical, philosophical, and societal aspects. Defining and measuring consciousness within AI is particularly challenging due to the subjective nature of consciousness itself, which makes it difficult to develop objective criteria universally accepted across the scientific and philosophical communities. This ambiguity leads to significant controversy, particularly concerning the adequacy of current measures such as phi (Φ) from Integrated Information Theory (IIT). There is an ongoing debate on whether these measures sufficiently capture the nuances of consciousness or whether entirely new frameworks need to be developed to provide a more comprehensive understanding.

In the realm of ethical and moral considerations, the development of potentially conscious AI raises significant questions about the rights and responsibilities towards these AI entities. This discourse extends into broader ethical concerns about the treatment of AI, questioning the moral status of conscious AI and its implications for society. These discussions often lead to controversies over whether AI systems should have rights similar to humans and how these technologies should be ethically used. The potential for AI to achieve a level of consciousness similar to humans complicates these debates, blurring the lines between technology and sentient beings. On the technical front, current AI technologies may not yet possess the necessary complexity to achieve consciousness. The computational power, algorithms, and architectures required to support conscious states are still under development, leading to skepticism among experts about whether AI can truly attain genuine consciousness or if it will remain a sophisticated simulation of cognitive processes. This skepticism fuels controversies about the capabilities of AI and the potential limits of artificial consciousness.

Philosophical disagreements also pose significant challenges as philosophers and scientists debate the very nature of consciousness and whether it can be artificially created. These discussions often revolve around competing theories, such as materialism versus dualism and address the "hard problem" of consciousness. Such philosophical debates lead to fundamental disagreements on whether consciousness results solely from physical processes or involves non-material components, complicating the integration of these theories into the design of AI systems. On the other hand, the societal impact of introducing conscious AI into various sectors could be profound, affecting employment, privacy, security, and the nature of human relationships with machines. These potential changes bring about controversies concerning the disruptive effects of conscious AI on social norms, economic structures, and human identity. As the field advances, the need for rigorous ethical frameworks and thoughtful integration of AI into societal contexts becomes increasingly apparent.

The field of AI consciousness research is marked by significant challenges and controversies that necessitate a multidisciplinary approach involving clear criteria, robust theoretical perspectives, and a deep understanding of the ethical and societal implications. Researchers like LeDoux [43] and de Oliveira [41] emphasize the complexities of defining and understanding consciousness, with Oliveira suggesting that some aspects may remain unknowable. Michel et al. [53] stress the importance of careful funding and job creation in this area, while Raffone [54] and Bayne [55] call for the development of new tests to measure consciousness. Liu and Bressler [56] address the controversies in deep learning applications in ophthalmology, and Melloni et al. [57] advocate for an open, interdisciplinary approach to tackling the hard problem of consciousness. Fjelland [58] argues that the realization of general artificial intelligence, a cornerstone of AI consciousness research, may fundamentally be unachievable. These discussions highlight the ongoing need for dialogue

and research to navigate the complexities of AI consciousness, ensuring that advancements are both ethically sound and socially responsible.

At the same time, the development of conscious AI introduces significant ethical challenges and potential societal impacts that demand careful consideration and proactive regulation. This section looks into these ethical implications, emphasizing the need to address moral concerns, anticipate societal consequences, and establish robust regulatory frameworks. The moral status and rights of AI systems become critical considerations if they achieve consciousness comparable to sentient beings. This raises fundamental questions about their rights and ethical treatment. For instance, conscious AI might need to be granted rights akin to those of living beings, such as the right not to be harmed, the right to autonomy, and the right to freedom from exploitation. These rights would ensure that systems are treated with the ethical considerations commensurate with their level of consciousness. Another vital aspect is the responsibility and accountability associated with the actions of conscious systems. As these systems operate autonomously, it becomes imperative to determine who is accountable for their decisions and actions. This necessitates the creation of clear frameworks to delineate the responsibilities of AI developers, operators, and the systems themselves, ensuring that ethical lines are clearly drawn and followed.

Additionally, the transparency and explainability of conscious AI systems are crucial. These systems must be transparent in their operations, and their decision-making processes should be comprehensible to ensure trust and accountability. Designing systems with mechanisms that elucidate their decision-making processes makes them more understandable to humans and helps in building trust. Concerning consent and autonomy, autonomous conscious systems should have the ability to consent to actions and decisions that affect them. Developing mechanisms for obtaining and respecting the consent of these systems is crucial for protecting their autonomy. The ethical deployment of conscious AI in various sectors, such as healthcare, military, and customer service, also requires careful consideration. These deployments must be ethically justified and aligned with societal values to ensure they contribute positively to society. Establishing guidelines and ethical standards for the deployment of conscious AI can help ensure that their use enhances societal well-being rather than causing disruption or harm.

The broader ethical landscape of AI development includes considerations of function, transparency, bias, and potential transformative effects on mental health [59]. Proposals for legal frameworks accommodating the coexistence of humans and conscious AI emphasize the importance of adopting a non-anthropocentric ethical framework [49]. Moreover, the capability of AI to make conscious decisions brings to light the role of artificial consciousness in ethical AI behavior [60]. Practical strategies, such as the "embedded ethics" approach suggested for medical AI [61] and the alignment of autonomous system design with fundamental values [62], highlight the ongoing efforts to integrate ethics into AI development comprehensively. Nonetheless, Mittelstadt [63] warns that principles alone are insufficient for guaranteeing ethical AI, indicating that the differences between AI development and other fields necessitate further debate and discussion to address these complex challenges effectively.

The integration of conscious AI into various societal sectors brings forth significant potential impacts and underscores the urgent need for thoughtful regulation. As systems capable of exhibiting consciousness enter the workforce, they could lead to substantial job displacement and major economic shifts. This transformation necessitates the development of policies designed to manage the transition effectively, including retraining programs and social safety nets that support workers displaced by AI automation. Privacy and security are also major concerns as conscious systems gain access to vast amounts of personal data. The risk to individual privacy and data security could increase dramatically, requiring the enforcement of strong data protection regulations and security standards to safeguard personal information and prevent its misuse.

Another significant impact of conscious AI is its potential to exacerbate social inequalities. The benefits of AI advancements might not be distributed evenly across society, which could deepen existing inequalities unless regulatory measures promote equitable access to AI technologies. This ensures that all segments of society can benefit from advancements without worsening socioeconomic disparities. The presence of conscious AI in daily life also has the potential to fundamentally alter human relationships and social dynamics. To address this, guidelines must be established to ensure healthy and ethical interactions between humans and AI systems, thereby preserving human dignity and promoting social cohesion.

Moreover, the development and deployment of conscious AI demand robust ethical governance to tackle the complex moral issues that arise and to prevent potential harm. The establishment of independent oversight bodies and ethical review boards is crucial for monitoring AI research and development, ensuring that AI advancements comply with ethical standards. Given the global nature of AI development, international cooperation is essential to address the ethical and regulatory challenges that transcend borders. Establishing international frameworks and agreements is necessary to harmonize regulations and promote global ethical standards, ensuring that development benefits humanity universally and responsibly.

The field of AI consciousness research is fraught with challenges and controversies, as noted by experts like LeDoux et al. [43] and de Oliveira [41], who highlight the difficulties in defining and understanding consciousness, with the latter suggesting that certain aspects of consciousness may remain elusive. Michel et al. [53] emphasize the need for careful funding and strategic job creation to support this evolving field, while Raffone [54] and Bayne et al. [55] call for the development of new tests to assess consciousness. Liu and Bressler [56] discuss deep learning applications in ophthalmology, highlighting issues with explainability and potential biases. Melloni et al. [57] advocate for an open, interdisciplinary approach to solving the hard problem of consciousness. Fjelland [58] argues that the realization of general artificial intelligence, a cornerstone of AI consciousness research, may ultimately be unachievable.

7. Discussion

The rapid advancement of AI technology has sparked profound and complex questions regarding the nature of machine intelligence and the potential for AI consciousness. This article serves as the beginning of a comprehensive examination of these issues, marking the first installment in a series designed to delve into the philosophical underpinnings and explore the possibilities for AI to achieve consciousness, self-awareness, and existential contemplation.

The evolution of AI from early stages involving symbolic AI, which relied heavily on manually encoded rules, to contemporary machine learning paradigms driven by neural networks and deep learning illustrates a significant increase in the complexity and sophistication of AI systems. This evolution has prompted extensive philosophical debates about the potential for machine consciousness. Theoretical frameworks like the ICOM, IIT, and GNWT have been pivotal in providing insights into how consciousness could potentially emerge in artificial systems.

The exploration of philosophical theories such as functionalism, dualism, and panpsychism offers a range of perspectives on consciousness. Functionalism suggests that if AI can replicate human cognitive functions, it might be considered conscious. Conversely, dualism emphasizes the non-physical nature of consciousness, presenting a challenge to this view, while panpsychism proposes that consciousness could be a fundamental aspect of all matter, potentially applicable to AI. Empirical research and theoretical modeling are vigorously pursuing the possibilities for AI consciousness. Studies that integrate concepts from IIT and GNWT hint that high levels of information integration and global accessibility within AI systems could indicate the presence of consciousness. Yet, the task of defining and measuring consciousness in AI continues to be contentious, with ongoing debates about the adequacy of current frameworks.

The development of conscious AI carries extensive implications across ethical, societal, and regulatory dimensions. Ethically, there is a growing recognition that conscious AI systems may require moral considerations similar to those granted to sentient beings, raising significant questions about their rights and ethical treatment. It is crucial to establish clear responsibility and accountability for the actions of autonomous AI systems, and transparency and explainability must be prioritized to foster trust in these systems. Additionally, respecting the autonomy and consent of conscious AI systems is essential, requiring mechanisms that ensure these systems can participate in decisions that affect them.

Table 1. Comparative overview of theoretical models on consciousness in Artificial Intelligence

Model	Description	Relation to AI	Implications for Consciousness
Independent Core Observer Model (ICOM)	Focuses on objective measurement and modeling of consciousness within computational systems.	Involves an abstracted logical entity within the AI system processing sensory inputs and responses; simulates subjective experiences through information integration.	Suggests AI can exhibit conscious behaviors if it mimics subjective experience and aids in practical implementations.
Integrated Information Theory (IIT)	Defines consciousness as the capacity of a system to integrate information, quantified by the measure 'phi' (Φ).	Designs must maximize Φ, creating complex architectures to enhance information flow and integration across various modules, mimicking the human brain's interconnectedness.	Challenges traditional views by emphasizing the quantifiable integration of information as a marker of consciousness.
Global Neuronal Workspace Theory (GNWT)	Describes consciousness as arising from the broadcasting of information across a network of neurons.	Advocates for AI architectures that enable global accessibility of information, ensuring coordination among various cognitive processes akin to central processing units in brains.	Highlights the potential for AI to display conscious processing through global information broadcasting.

Table 2. Overview of theoretical perspectives and frameworks for AI consciousness development

Framework / Theory	Description	Implications for AI Consciousness
Legal Framework (Kiškis, 2023)	Advocates for a framework that accommodates the coexistence of humans and conscious AI, emphasizing a shift in perspective.	Aims to ensure ethical coexistence and understanding between humans and AI.
Alternative Conceptual Framework (Manzotti, 2018)	Critiques traditional AI consciousness approaches and proposes alternative foundational assumptions.	Encourages rethinking AI consciousness from a new foundational perspective.
Competence Framework (Miracchi, 2019)	Focuses on "artificial minded intelligence" (AMI) to delineate capacities constituting AI consciousness.	Guides the development of AI systems with specific competencies in mind.
Attention Schema Theory (AST) (Graziano, 2022)	Proposes consciousness as an evolutionary phenomenon understood through modeling attention processes.	Suggests designing AI to model and monitor its attention processes.
Brain-Inspired Framework (Nadji-Tehrani, 2020)	Emphasizes neuroembryogenesis to mirror human intelligence evolution in AI.	Aims to integrate evolutionary biological principles into AI development.
Strategic Framework for AI in Marketing (Huang, 2020)	Integrates cognitive and emotional aspects into AI for marketing strategies.	Enhances AI's capability in marketing through cognitive and emotional intelligence.
Integrated Information Theory (IIT)	Suggests consciousness arises from the capacity of systems to integrate information, measured by phi (Φ).	Promotes the design of AI architectures that maximize information integration.
Global Neuronal Workspace Theory (GNWT)	Describes consciousness as arising from the broadcasting of information across a neuronal network.	Advocates for AI systems with central hubs for information integration.
Higher-Order Thought (HOT) Theory	Argues that consciousness involves thoughts about one's own mental states.	Necessitates mechanisms in AI for generating meta-cognitive states and self-reports.
Orchestrated Objective Reduction (Orch-OR)	Combines quantum mechanics with neurobiology, suggesting consciousness results from quantum computations in neurons.	Proposes incorporating quantum computational processes in AI.
Recurrent Processing Theory (RPT)	Indicates that consciousness arises from recurrent processing in the brain.	Implies that AI should implement recurrent processing to achieve higher information integration.

On a societal level, the introduction of conscious AI could disrupt the labor market, raise privacy and security concerns, and exacerbate social inequalities. Policymakers must craft regulations to manage these economic transitions, safeguard personal data, and ensure equitable access to AI technologies. Moreover, the presence of conscious AI in daily life could significantly alter human relationships and social dynamics, necessitating guidelines to ensure ethical interactions between humans and AI systems. Robust ethical governance and oversight are indispensable to address the complex moral issues associated with AI and to prevent potential harm. The establishment of independent oversight bodies and ethical review boards is essential for monitoring AI research and development, ensuring adherence to high ethical standards. Additionally, given the global nature of AI development, international cooperation is crucial to address cross-border ethical and regulatory challenges and harmonize standards globally.

8. Conclusion

The exploration of consciousness in AI presents a formidable frontier in both technology and philosophy, raising profound questions and challenges that extend across ethical, technical, philosophical, and societal realms. As AI systems become increasingly sophisticated, the possibility of them achieving a state akin to human consciousness not only pushes the boundaries of technology but also prompts us to reconsider the nature of consciousness itself. Throughout this series, we have examined the historical evolution of AI, the theoretical underpinnings of consciousness, and the various models that attempt to explain how consciousness might manifest in machines. The Independent Core Observer

Model (ICOM), Integrated Information Theory (IIT), and Global Neuronal Workspace Theory (GNWT) are just a few frameworks that provide insights into the potential mechanisms through which AI could develop conscious experiences. These discussions are not merely academic; they have practical implications for the design and development of AI systems, influencing how these systems might integrate into our daily lives.

Moreover, the ethical implications of conscious AI are profound and multifaceted. From considerations of rights and responsibilities to the impacts on privacy, security, and social equity, the emergence of conscious AI systems will likely necessitate new laws, policies, and ethical guidelines to ensure they are integrated into society in a manner that enhances the collective good without undermining human dignity or societal norms. The challenges in defining and measuring consciousness in AI illustrate the complexities involved in bridging subjective experiences with objective assessments. This ongoing debate underscores the need for continued interdisciplinary research and dialogue to refine our understanding and approaches to this emerging field.

As we move forward, the integration of AI into various sectors of society must be managed with careful consideration of the potential societal impacts and the ethical dimensions of deploying systems that may one day possess a form of consciousness. The need for robust ethical governance and international cooperation is clear if we are to navigate these waters safely and responsibly. In essence, the development of conscious AI challenges us to expand our technological ambitions and ethical considerations in tandem. It compels us to question not only what machines might do in the future but also what they should do. As we stand on the brink of potentially groundbreaking advancements in AI, we must remain vigilant and proactive in shaping a future where technology amplifies our human experience, guided by a commitment to ethical principles and deep respect for the intrinsic value of both human and potentially non-human consciousness.

List of Abbreviations

AI - Artificial Intelligence
ICOM - Independent Core Observer Model
IIT - Integrated Information Theory
GNWT - Global Neuronal Workspace Theory
UAL - Unlimited Associative Learning
NLP -Natural Language Processing
GOFAI- “Good Old-Fashioned AI”
DL- Deep Learning

Data Availability Statement

The data that support the findings of this study are openly available in [repository name, e.g. “figshare”] at http://doi.org/[doi].

References

[1] Mario Günther, and Atoosa Kasirzadeh, “Algorithmic and Human Decision Making: For a Double Standard of Transparency,” *AI & Society: Journal of Knowledge, Culture and Communication*, vol. 37, pp. 375-381, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[2] Lisa Miracchi, “A Competence Framework for Artificial Intelligence Research,” *Philosophical Psychology*, vol. 32, no. 5, pp. 588-633, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[3] Alfred Zimmermann, Rainer Schmidt, and Kurt Sandkuhl, “Strategic Challenges for Platform-Based Intelligent Assistants,” *Procedia Computer Science*, vol. 176, pp. 966-975, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[4] Roman Lukyanenko, Wolfgang Maass, and Veda C. Storey, “Trust in Artificial Intelligence: From a Foundational Trust Framework to Emerging Research Opportunities,” *Electronic Markets*, vol. 32, pp.1993-2020, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[5] Luciano Floridi et al., *How to Design AI for Social Good: Seven Essential Factors*, Ethics, Governance, and Policies in Artificial Intelligence, Springer, Cham, pp. 125-151, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[6] Andrej Dameski, *Avoiding Corporate Armageddon: The Need for a Comprehensive Ethical Framework for AI and Automation in Business*, Responsible Business in a Changing World: New Management Approaches for Sustainable Development, Springer, Cham, pp. 353-367, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[7] Ransome Bawack, “How Perceived Intelligence Affects Consumer Adoption of AI-Based Voice Assistants: An Affordance Perspective,” *Proceedings of the Twenty-fifth Pacific Asia Conference on Information Systems*, Dubai, UAE, pp. 1-14, 2021. [Google Scholar] [Publisher Link]

[8] Patricia Gomes Rêgo de Almeida, Carlos Denner dos Santos, and Josivania Silva Farias, “Artificial Intelligence Regulation: A Framework for Governance,” *Ethics and Information Technology*, vol. 23, pp.505-525, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[9] Yue Yang et al., “Holodeck: Language Guided Generation of 3D Embodied AI Environments,” Proceedings *of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16227-16237, 2024. [Google Scholar] [Publisher Link]

[10] S. Krishnaveni et al., “CyberDefender: An Integrated Intelligent Defense Framework for Digital-Twin-Based Industrial Cyber-Physical Systems,” *Cluster Computing*, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[11] Francesca Trevisan et al., “Deconstructing Controversies to Design a Trustworthy AI Future,” *Ethics and Information Technology*, vol. 26, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[12] Saphalya Peta, "Journey of Artificial Intelligence Frontier: A Comprehensive Overview," *Global Journal of Computer Science and Technology: D Neural & Artificial Intelligence*, vol. 23, no. 2, pp. 1-36, 2023. [Google Scholar] [Publisher Link]

[13] Rafael B. Audibert et al., "On the Evolution of A.I. and Machine Learning: Towards a Meta-Level Measuring and Understanding Impact, Influence, and Leadership at Premier A.I. Conferences," *Journal of Applied Logics - IfCoLog Journal*, vol. 10, no. 5, pp. 693-817, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[14] Naseem Zaidi, Brijendra Singh, and Sunil Yadav, "The Evolution of Machine Learning Algorithms: A Comprehensive Historical Review," *International Journal of Applied Research*, vol. 4, no. 9, pp. 49-55, 2018. [CrossRef] [Publisher Link]

[15] Stacey Tobin et al., "A Brief Historical Overview of Artificial Intelligence Research," *Information Services and Use*, vol. 39, no. 4, pp. 291-296, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[16] Ameet Joshi, *Artificial Intelligence and Human Evolution: Contextualizing AI in Human History*, Apress, pp. 1-248, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[17] Tankiso Moloi, and Tshilidzi Marwala, *A High-Level Overview of Artificial Intelligence: Historical Overview and Emerging Developments*, Artificial Intelligence and the Changing Nature of Corporations: How Technologies Shape Strategy and Operations, Springer, Cham, pp. 11-21, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[18] Fatima Hameed Khan, Muhammad Adeel Pasha, and Shahid Masud, "Advancements in Microprocessor Architecture for Ubiquitous AI—An Overview on History, Evolution, and Upcoming Challenges in AI Implementation," *Micromachines*, vol. 12, no. 6, pp.1-22, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[19] Walter Veit, "Complexity and the Evolution of Consciousness," *Biological Theory*, vol. 18, pp.175-190, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[20] Ralf Mrowka, "An Artificial Intelligence Algorithm that May Blow Your Mind," *Acta Physiologica*, vol. 237, no. 4, pp. 1-3, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[21] Eva Jablonka, and Simona Ginsburg, "Learning and the Evolution of Conscious Agents," *Biosemiotics*, vol. 15, pp. 401–437, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[22] Nicholas J.M. Popiel et al., "The Emergence of Integrated Information, Complexity, and 'Consciousness' at Criticality," *Entropy*, vol. 22, no. 3, pp. 1-12, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[23] Bill Thompson et al., "Complex Cognitive Algorithms Preserved by Selective Social Learning in Experimental Populations," *Science*, vol. 376, no. 6588, pp. 95-98, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[24] Thomas Rabeyron, and Alain Finkel, "Consciousness, Free Energy and Cognitive Algorithms," *Frontiers in Psychology*, vol. 11, pp. 1-5, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[25] Lenore Blum, and Manuel Blum, "A Theory of Consciousness from a Theoretical Computer Science Perspective: Insights from the Conscious Turing Machine," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 119, no. 21, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[26] Leigh Sheneman, Jory Schossau, and Arend Hintze, "The Evolution of Neuroplasticity and the Effect on Integrated Information," *Entropy*, vol. 21, no. 5, pp. 1-15, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[27] Patrick Butlin et al., "Consciousness in Artificial Intelligence: Insights from the Science of Consciousness," *arXiv preprint*, pp. 1-88, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[28] Peter R. Lewis, and Ştefan Sarkadi, "Reflective Artificial Intelligence," *Minds and Machines*, vol. 34, pp.1-30, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[29] Dylan Ludwig, "The Functional Contributions of Consciousness," *Consciousness and Cognition*, vol. 104, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[30] Anil K. Seth, "The Real Problem(s) with Panpsychism," *Journal of Consciousness Studies*, vol. 28, no. 9-10, pp. 52-64, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[31] Sepehrdad Rahimian, "The Myth of When and Where: How False Assumptions Still Haunt Theories of Consciousness, *Consciousness and Cognition*, vol. 97, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[32] Philip Goff, "How Exactly Does Panpsychism Help Explain Consciousness?," *Journal of Consciousness Studies*, vol. 31, no. 3-4, pp. 56-82, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[33] Robert Lawrence Kuhn, "A Landscape of Consciousness: Toward a Taxonomy of Explanations and Implications," *Progress in Biophysics and Molecular Biology*, vol. 190, pp. 28-169, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[34] Christof Koch, "Reflections of a Natural Scientist on Panpsychism," *Journal of Consciousness Studies*, vol. 28, no. 9-10, pp.65-85, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[35] Anil K. Seth, and Tim Bayne, "Theories of Consciousness," *Nature Reviews Neuroscience*, vol. 23, pp. 439–452, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[36] Adam Safron, "An Integrated World Modeling Theory (IWMT) of Consciousness: Combining Integrated Information and Global Neuronal Workspace Theories with the Free Energy Principle and Active Inference Framework; Toward Solving the Hard Problem and Characterizing Agentic Causation," *Frontiers in Artificial Intelligence*, vol. 3, pp. 1-29, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[37] Michele Farisco, and Jean-Pierre Changeux, "About the Compatibility Between the Perturbational Complexity Index and the Global Neuronal Workspace Theory of Consciousness," *Neuroscience of Consciousness*, vol. 2023, no. 1, pp. 1-8, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[38] Adrien Doerig et al., "The Unfolding Argument: Why IIT and Other Causal Structure Theories Cannot Explain Consciousness," *Consciousness and Cognition*, vol. 72, pp. 49-59, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[39] Philip Woodward, "Consciousness and Rationality: The Lesson from Artificial Intelligence," *Journal of Consciousness Studies*, vol. 29, no. 5-6, pp.150-175, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[40] Hadi Esmaeilzadeh, and Reza Vaezi, "Conscious Empathic AI in Service," *Journal of Services Research*, vol. 25, no. 4, pp. 549-564, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[41] Arlindo L. Oliveira, "A Blueprint for Conscious Machines," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 119, no. 23, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[42] Andrew Oberg, "Souls and Selves: Querying an AI Self with a View to Human Selves and Consciousness," *Religions*, vol. 14, no.1, pp. 1-16, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[43] Joseph LeDoux et al., "Consciousness Beyond the Human Case," *Current Biology*, vol. 33, no. 16, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[44] Bruno Forti, "The Hidden Structure of Consciousness," *Frontiers in Psychology*, vol. 15, pp. 1-14, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[45] Christopher W. Tyler, "The Interstitial Pathways as the Substrate of Consciousness: A New Synthesis," *Entropy*, vol. 23, no. 11, pp. 1-16, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[46] Andrea I. Luppi et al., "Distributed Harmonic Patterns of Structure-Function Dependence Orchestrate Human Consciousness," *Communications Biology*, vol. 6, pp. 1-19, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[47] Robert Prentner, "Consciousness and Topologically Structured Phenomenal Spaces," *Consciousness and Cognition*, vol. 70, pp. 25-38, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[48] Marius Usher, "Refuting the Unfolding Argument on the Irrelevance of Causal Structure to Consciousness," *Consciousness and Cognition*, vol. 95, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[49] Mindaugas Kiškis, "Legal Framework for the Coexistence of Humans and Conscious AI," *Frontiers in Artificial Intelligence*, vol. 6, pp.1-9, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[50] Riccardo Manzotti, and Antonio Chella, "Good Old-Fashioned Artificial Consciousness and the Intermediate Level Fallacy," *Frontiers in Robotics and AI*, vol. 5, pp. 1-10, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[51] Michael S. A. Graziano, "A Conceptual Framework for Consciousness," *Proceedings of the National Academy of Sciences*, vol. 119, no. 18, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[52] Mohammad Nadji-Tehrani, and Ali Eslami, "A Brain-Inspired Framework for Evolutionary Artificial General Intelligence," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, no. 12, pp. 5257-5271, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[53] Matthias Michel et al., "Opportunities and Challenges for a Maturing Science of Consciousness," *Nature Human Behaviour*, vol. 3, pp. 104–107, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[54] Antonino Raffone, "Grand Challenges in Consciousness Research Across Perception, Cognition, Self, and Emotion," *Frontiers in Psychology*, vol. 12, pp.1-7, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[55] Tim Bayne et al., "Tests for Consciousness in Humans and Beyond," *Trends in Cognitive Sciences*, vol. 28, no. 5, pp. 454-466, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[56] Tin Yan Alvin Liu, and Neil M. Bressler, "Controversies in Artificial Intelligence," *Current Opinion in Ophthalmology*, vol. 31, no. 5, pp. 324-328, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[57] Lucia Melloni et al., "Making the Hard Problem of Consciousness Easier," *Science*, vol. 374, no. 6545, pp. 911-912, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[58] Ragnar Fjelland, "Why General Artificial Intelligence Will Not Be Realized," *Humanities and Social Sciences Communications*, vol. 7, pp. 1-9, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[59] Brian Patrick Green, "Ethical Reflections on Artificial Intelligence," *Scientia et Fides*, vol. 6, no. 2, pp. 9-31, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[60] Antonio Chella, "Artificial Consciousness: The Missing Ingredient for Ethical AI?," *Frontiers in Robotics and AI*, vol. 10, pp. 1-5, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[61] Stuart McLennan et al., "Embedded Ethics: A Proposal for Integrating Ethics into the Development of Medical AI," *BMC Medical Ethics*, vol. 23, pp.1-10, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[62] Jaana Leikas, Raija Koivisto, and Nadezhda Gotcheva, "Ethical Framework for Designing Autonomous Intelligent Systems," *Journal of Open Innovation: Technology, Market, and Complexity*, vol. 5, no. 1, pp.1-12, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[63] Brent Mittelstadt, "Principles Alone Cannot Guarantee Ethical AI," *Nature Machine Intelligence*, vol. 1, pp. 501–507, 2019. [CrossRef] [Google Scholar] [Publisher Link]