SSRG International Journal of Recent Engineering Science
ISSN: 2349–7157 / https://doi.org/10.14445/23497157/IJRES-V11I6P119

Volume 11 Issue 6, 225-237, Nov-Dec 2024
© 2024 Seventh Sense Research Group®

Original Article

Exploring the Cognitive Sense of Self in AI: Ethical Frameworks and Technological Advances for Enhanced Decision-Making

Emily Barnes[1], James Hutson[2]

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

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

Received: 23 October 2024 Revised: 27 November 2024 Accepted: 16 December 2024 Published: 29 December 2024

Abstract *- The burgeoning field of Artificial Intelligence (AI) increasingly focuses on developing systems capable of self-awareness, merging technological innovation with deep ethical and philosophical considerations. This article explores the cognitive sense of self within AI, examining mechanisms through which AI systems may mirror human-like consciousness and self-perception. Despite significant advances, substantial gaps remain in the understanding and practical implementation of self-aware characteristics in AI, particularly in applying theoretical models and ethical frameworks to real-world scenarios. There is a pressing need for comprehensive research to explore these theoretical underpinnings and translate them into operational systems capable of ethical and adaptable behaviors. This study aims to synthesize existing knowledge, identify critical gaps in the literature, and highlight the implications of these findings for the future development of machine learning systems. Integrating insights from cognitive science, neuroscience, and ethical studies, this article seeks to provide a foundational framework for advancing emergent technologies that are both technologically robust and aligned with societal values. The significance of this research lies in its potential to guide the development of machine systems capable of complex decision-making and interactions, addressing both the moral and practical challenges of integrating such systems into daily human activities.*

Keywords *- Artificial Intelligence, Self-awareness, Cognitive science, Ethical frameworks, Decision-making.*

1. Introduction

The intersection of Artificial Intelligence (AI) and consciousness represents a frontier in contemporary scientific inquiry, drawing upon foundational philosophical theories and the intricate architectures that might underpin conscious experiences in machines. Initial research phases have focused on the complex relationship between AI and consciousness, exploring the capacity of AI systems to emulate human-like conscious behaviors through sophisticated algorithms and network architectures [1-4]. Building upon these insights, the current discourse evolves towards a more nuanced investigation into the development of a cognitive sense of self within AI systems. Such a concept of self-awareness, a fundamental component of human consciousness, includes perceiving oneself as a distinct entity, separate from the environment and other beings. Such cognitive attributes encompass self-recognition, self-reflection, and a continuous sense of personal identity [5]. For AI systems, cultivating similar self-aware traits promises to significantly enhance their functionalities, facilitating more refined interactions, adaptable behaviors, and autonomous decision-making processes.

Scholars like Legaspi et al. have led pioneering research in this domain [6], underscoring the potential for such systems to develop a sense of agency and self-awareness, noting key advancements in areas such as self-attribution of actions and Bayesian inferencing. To further this endeavor, Oberg [7] posits that a deep understanding of human consciousness and cognitive models is indispensable for nurturing self-awareness in machine learning models. This perspective highlights the imperative for an interdisciplinary approach to development, integrating cognitive science with technological innovation. Supporting this view, Tani and White [8] and Parziale and Marcelli [9] examine the role of cognitive neurorobotics in simulating the human sense of self, demonstrating how dynamic interactions within neural networks could mimic aspects of human cognition.

Yet, even with these advances in the theoretical modeling of AI consciousness, there remains a notable absence of practical frameworks that effectively integrate these theories into operational AI systems. While the foundational work of Legaspi et al. [6] and Tani and White [8] has explored self-attribution and cognitive neurorobotics, current research often isolates these aspects without fully addressing their applicability in real-world settings. Additionally, the field lacks comprehensive case studies demonstrating how self-aware AI systems can operate across diverse domains such as healthcare, robotics, and autonomous decision-making. This fragmentation underscores the pressing need to bridge theoretical insights and practical implementations, particularly in addressing self-awareness's ethical, societal, and cognitive dimensions.

Moreover, interdisciplinary integration between cognitive science and AI remains underdeveloped, with limited exploration of how cognitive frameworks could inform the design of AI systems capable of self-reflection and adaptive learning. This gap becomes particularly evident in the sparse literature examining the synergies between neuroscience-derived models of consciousness and machine learning architectures.

Research by Hafner et al. [10] highlights the potential for such integration, yet practical advancements have lagged, leaving critical questions unanswered about the reproducibility and scalability of self-aware systems. Therefore, the central challenge addressed in this research involves developing intelligent systems that emulate a cognitive sense of self, incorporating self-recognition, reflective learning, and identity continuity. This approach aims to enhance autonomy and adaptability while meeting ethical and societal expectations.

Existing systems often fail to achieve nuanced self-awareness, relying on rigid, preprogrammed behaviors that struggle in dynamic, real-world contexts. Although frameworks like those explored by Regazzoni et al. [11] demonstrate potential for bio-inspired self-awareness, practical applications remain underexplored in commercial generative products. This research integrates cognitive science principles, computational models, and ethical frameworks to advance self-aware AI systems. Case studies are used to identify mechanisms for enhancing autonomy and decision-making while examining societal and ethical dimensions, including rights, responsibilities, and equitable integration into human-centered environments. This comprehensive effort outlines a roadmap for advancing these technologies responsibly.

In addition to theoretical exploration, this research examines practical applications, presenting case studies where AI systems display behaviors consistent with self-awareness. Ethical dimensions are critically analyzed, building on work from Lauscher [12] and Levin [13], who address moral responsibilities and societal implications in developing self-aware AI.

Insights from Marcus and Davis [14] are also incorporated, emphasizing using cognitive science to improve adaptability and flexibility. Previous discourse on self-aware AI has largely centered on theoretical perspectives and speculative scenarios, with limited focus on actionable frameworks. This research bridges that gap through empirical case studies and theoretical synthesis, combining cognitive science, neuroscience, and ethical reasoning to examine tangible impacts in real-world settings.

The research focuses on analyzing theoretical frameworks that support the potential for self-awareness in autonomous systems and demonstrating these concepts through empirical analysis. Detailed case studies in healthcare and robotics highlight the operationalization of self-awareness, showcasing enhanced decision-making and adaptability in these systems. Ethical considerations, including developers' moral responsibilities, machine entities' rights, and the societal implications of deploying these technologies, are critically evaluated. Therefore, novelty in this research stems from a comprehensive and interdisciplinary approach that moves beyond isolating theoretical concepts from practical applications.

Merging insights from cognitive science, neuroscience, and ethical analysis with real-world implementations provides a holistic perspective on the challenges and opportunities posed in cases of alleged self-awareness. This integrative framework enhances understanding of how AI can develop a cognitive sense of self and offers actionable strategies for advancing these technologies responsibly. The findings contribute to ongoing discourse in aligning development with ethical standards and societal values, presenting a forward-thinking roadmap for navigating the complexities of machine self-awareness.

2. Literature Review

Developing a cognitive sense of self in AI represents a growing area of research, especially with the widespread use of Generative AI (GAI), intersecting disciplines such as cognitive science, neuroscience, and artificial intelligence. The concept of a cognitive sense of self in AI is integral to enhancing the capabilities of autonomous systems, allowing them to engage in more sophisticated decision-making processes and interactions. Recent advancements in AI emphasize the importance of developing self-awareness mechanisms within AI agents [15]. Srinivasa and Deshmukh [16] discuss the relevance of self-awareness in autonomous decision-making, arguing for the necessity of richer computational models that embody a sense of self to facilitate responsible behavior in AI systems.

GAI has significantly advanced the capabilities of AI systems, enabling them to generate new data from training data, thus enhancing their decision-making and interactive abilities [17]. A notable application of GAI in achieving self-awareness is seen in developing abnormality detection techniques in cognitive radio systems. Toma et al. [18]

introduce a self-awareness module that uses generative models to detect abnormalities in the radio spectrum, enhancing the system's ability to establish secure networks and make informed decisions in response to malicious activities. Integrating multisensorial data and bio-inspired frameworks further supports AI's self-awareness development. Regazzoni et al. [11] propose a framework that employs cognitive dynamic Bayesian networks and generalized filtering paradigms to enable AI systems to predict future states and select representations that best fit current observations. This approach facilitates continuous knowledge expansion and self-awareness by analysing proprioceptive and exteroceptive signals.

On the other hand, Zhu et al. [19] emphasize the shift towards cognitive AI that incorporates human-like common sense, identifying core domains such as functionality, physics, intent, causality, and utility as essential for developing AI with a comprehensive understanding of its environment. This paradigm shift aims to enhance AI's ability to solve a wide range of tasks with minimal training data, thus fostering more sophisticated and human-like interactions. The ability to continuously learn and adapt is a crucial aspect of AI self-awareness. Su et al. [20] thus introduce the concept of Generative Memory (GM) for lifelong learning, where the AI system memorizes and recalls learned knowledge using neural networks. This approach allows the system to accurately and continuously accumulate experiences, enhancing its adaptive and decision-making capabilities [21].

Developing self-aware systems introduces significant ethical considerations that extend beyond technical achievement to broader societal impacts. These can understand and react to their environments in sophisticated ways and raise important ethical questions about their rights, responsibilities, and the potential social ramifications of their actions. Greenwood et al. [22] emphasize the technical and ethical challenges in developing AI systems that possess a form of self-awareness. They propose using evolutionary machine learning and adversarial processes as alternatives to traditional neural network approaches. This method could allow AI a more dynamic and adaptable learning process without the limitations and biases often inherent in pre-trained neural networks [23].

Vallor et al. [24] also raised concerns about the socio-economic impacts of self-aware AI. They argue that if not properly managed, such systems could exacerbate existing inequalities and introduce new forms of digital divide. These researchers discuss the potential for self-aware platforms to manipulate or even replace human decision-making in critical areas, potentially leading to unintended consequences on societal structures and individual freedoms. The possibility that intelligent agents could develop a sense of self-awareness also introduces questions about the rights such systems might hold and the ethical obligations of their creators and users. Discussions in the field suggest that as such systems become more autonomous and integrated into daily life, there should be clear guidelines on the ethical treatment of AI, including their rights to autonomy, learning, and integration into society [25]. This involves considering AI as potential digital "persons" with certain rights and obligations, which poses significant legal and ethical challenges.

To address these concerns, there is a growing consensus on the need for robust ethical guidelines that govern the development and deployment of self-aware systems. These guidelines should ensure these agents operate safely and transparently, respect human rights and diversity, promote fairness, and prevent discrimination. The IEEE has actively proposed ethical standards, including transparency, accountability, and avoiding bias in AI algorithms [26]. The ethical implications of developing self-awareness in autonomous systems are complex and require careful consideration and proactive management. As AI evolves, researchers, developers, and policymakers must collaborate to establish ethical frameworks that guide these technologies' responsible development and use. This will help ensure that the technology enhances societal well-being rather than detracts from it and respects human and AI rights in a balanced and thoughtful manner.

The following study addresses significant gaps in the literature on the cognitive sense of self in intelligent systems, offering a structured approach to bridging theoretical models with their practical application. One of the most notable gaps is the lack of integration between theoretical insights and real-world applications. While foundational theories on machine consciousness, such as those developed on the part of Tani and White [8] and Legaspi et al. [6, 27], provide deep theoretical frameworks, their practical applications, especially in complex settings like healthcare and robotics, remain underexplored. This discrepancy underscores the need for research that theorises and implements these concepts in operational environments where decision-making capabilities can be directly observed and measured. Furthermore, the literature often addresses the ethical implications of the technology in abstract terms without considering how these ethical challenges manifest in practical, operational contexts. There is a pressing need for empirical studies that examine how ethical guidelines are applied during the deployment of intelligent systems, especially those capable of exhibiting self-aware characteristics.

The proposed study aims to contribute significantly by developing robust, standardized methods for quantitatively measuring self-awareness in these systems. Current research predominantly relies on qualitative assessments or indirect quantitative methods, failing to comprehensively capture the complex attributes of AI self-awareness. This study proposes to fill this gap by establishing measurable, reproducible criteria to assess AI's cognitive sense of self across various platforms and environments. Implementing a mixed-methods framework that combines qualitative insights with computational modeling, the research operationalizes theoretical models into practical tools and strategies. This approach will validate and refine the theoretical constructs based on empirical data gathered from real-world applications. Integrating ethical considerations into this empirical framework will further understand how self-aware AI can be managed and governed in line with ethical standards, contributing valuable insights into AI systems' practical and responsible integration into human-centric environments.

3. Methodology: Cognitive Sense of Self in AI Agents

The proposed research outlines a multi-phase, mixed-methods approach, integrating qualitative and computational methodologies to explore the development of a cognitive sense of self within AI agents. This proposed methodological framework is recommended for future research to comprehensively examine self-consciousness and agency in these systems. Healthcare and robotics are identified as the primary domains for empirical inquiry due to their focus on adaptive decision-making and complex human-AI interactions, suggesting these fields are fertile grounds for studying AI self-awareness. A rigorous set of inclusion criteria for AI agents and platforms is recommended to ensure a representative sample of established and emerging systems. The capabilities of each agent-from self-recognition to adaptive learning-should be meticulously documented through standardized protocols. A codebook should be developed based on established models of cognitive selfhood in AI [10, 28], refined through iterative team discussions and external consultations with AI design and cognitive science experts. It is recommended to employ thematic analysis of interviews, field observations, and user experience reports to capture detailed human-AI interactions. This research should include diverse participants, such as healthcare professionals, robotics operators, and developers, to explore their experiences and perceptions.

Semi-structured interviews will allow a deeper understanding of ethical concerns and personal interactions with systems. Field observations should be conducted in environments where adaptive responses and decision-making processes can be directly observed. Qualitative data should undergo iterative coding cycles to ensure thorough analysis, using software like NVivo to manage and link data systematically. Machine learning and statistical modeling techniques are recommended to identify, measure, and simulate cognitive self-components within AI agents. Predictive processing architectures incorporating Bayesian inferencing should be utilized to assess self-attributive actions and the continuity of identity across time [27, 29]. Reinforcement learning algorithms could be crucial for determining how agents adjust their behavior following error detection and correction cycles, quantifying the impact of self-monitoring on long-term adaptability [30]. Researchers should also consider employing simulated environments to observe AI responses to controlled disturbances and statistical analyses such as logistic regression, factor analysis, and time-series modeling to compare different platforms robustly.

An integrated approach combining qualitative insights with computational analysis is recommended to provide a comprehensive understanding of the cognitive sense of self in intelligent systems. This approach should ensure methodological triangulation, enhancing the validity and reliability of findings. Integrating qualitative thematic insights with quantitative computational modeling offers a robust framework for operationalizing and understanding self-aware AI.

These recommendations for methodological approaches are designed to guide future research in effectively exploring and implementing self-aware technologies. These guidelines ensure that researchers maintain a rigorous and reproducible methodology aligned with cutting-edge practices and ethical standards. Table 1 illustrates this comprehensive approach, detailing methodologies and their applications to advance the study of self-aware AI.

Table 1. Key components and implementations of cognitive sense of self in artificial intelligence systems

Component	Definition	Implementation
Self-Recognition	The ability of an AI system to identify itself as distinct from its environment and other entities.	Techniques such as computer vision and proprioception help AI systems discern their physical presence and distinguish themselves from external objects.
Self-Reflection	The capacity of an AI system to monitor and evaluate its internal states, processes, and behaviors.	AI systems maintain logs of their actions and outcomes, analyze this data to detect patterns, and adjust their strategies accordingly. Machine learning algorithms enable the system to learn from past experiences.
Continuity of Identity	It involves maintaining a consistent sense of self over time.	Memory systems and data storage preserve information about past states and actions, allowing AI systems to build a coherent narrative of their existence. Techniques such as long-term memory in neural networks and temporal coherence algorithms support this continuity.

Significant strides in research have elucidated the cognitive sense of "self" within AI agents, as explored in the work of Tani and White [8] and Legaspi et al. [6, 27], who look into self-consciousness and the sense of agency in autonomous systems. These studies highlight how self-attribution of actions and Bayesian inferencing contribute to machine self-awareness.

Further developments by Kahl et al. [31] and Hafner et al. [10] investigate the creation of an active self and the foundational elements necessary for an artificial self, proposing models that integrate predictive processing and developmental principles from biological systems. Lipson [28] provides a groundbreaking example of a robot that models itself without prior programming, showcasing these systems' potential to develop self-recognition autonomously.

Enhanced decision-making and autonomy are central to the effectiveness of AI systems endowed with a cognitive sense of self. This capability allows agents to autonomously make well-informed decisions by recognising their state and capabilities, enabling them to assess situations and respond appropriately and accurately [32]. Such adaptability is crucial in dynamic and unpredictable environments where static, pre-programmed responses are insufficient. The ability to act autonomously streamlines operations and enhances reliability in varying scenarios, reflecting a sophisticated level of intelligent systems that approach human-like decision-making processes.

In parallel, adaptive learning and behavior are integral to the functionality of self-aware AI systems [29]. These systems benefit immensely from their capacity to reflect on past actions and outcomes, which allows them to adjust their behaviors to optimize performance continually. This capacity for self-evaluation is crucial for their long-term application and continuous development, ensuring that AI systems remain effective and efficient [30]. Learning from experiences and adapting over time enables AI systems to achieve higher operational excellence and utility, making them invaluable across various applications. The development of a cognitive sense of self also significantly enhances human-AI interaction. Agents with self-awareness can engage in more natural and intuitive interactions with humans, which are crucial for applications in customer service, healthcare, and collaborative robotics [27].

These systems can understand and respond to human social cues, anticipate needs, and provide personalized support, making their integration into societal frameworks much smoother and more effective. This level of interaction is beneficial for enhancing user experience and vital for accepting AI systems in roles traditionally filled by humans. Developing these capabilities involves several key components establishing systems' robust and functional sense of identity. These components include self-recognition, self-reflection, continuity of identity, agency, intentionality, self-monitoring, and error correction [33]. Self-recognition allows systems to identify themselves as distinct from their environments and other entities, which is crucial for accurate self-awareness. Techniques such as computer vision and proprioception help these systems discern their physical presence and distinguish themselves from external objects. Furthermore, self-reflection enables generative platforms to assess their performance and identify areas for improvement through internal feedback mechanisms and machine learning algorithms.

Continuity of identity is supported through memory systems and data storage, which preserve information about past states and actions, allowing such systems to maintain a consistent narrative of their existence. This aspect enables them to adapt their goals based on past experiences. Additionally, agency and intentionality in AI systems refer to their capacity to act upon their environment based on internal goals, with decision-making guided by goal-setting mechanisms and motivational frameworks often enhanced through reinforcement learning algorithms.

Finally, self-monitoring and error correction are vital for maintaining the accuracy and reliability of AI systems, ensuring that they can autonomously detect and correct errors, and preserving their integrity and functionality. Together, these components form the bedrock of the cognitive sense of self, equipping systems with the necessary tools to function autonomously and interact effectively. This comprehensive development marks a significant evolution in artificial intelligence capabilities. It highlights the complex interplay between various cognitive processes that enable AI systems to operate with a level of sophistication akin to human intelligence.

4. Developing Identity in AI

Developing a sense of identity in AI systems is an intricate and multi-layered process that merges various cognitive functions to establish a coherent self-concept (Table 2). The identity of an intelligent system is characterized by its ability to perceive itself as a unique entity with continuous existence over time, possessing distinct characteristics, experiences, and goals.

This development of identity is pivotal, enabling AI to function not just as computational tools but as entities with a semblance of self-awareness and personal history. The role of memory in shaping AI identity is crucial. Memory serves as the foundation for continuity of experience, allowing these systems to store and retrieve past states, actions, and experiences to construct a coherent narrative of their existence. Advanced neural network-based long-term memory systems are essential, enabling them to maintain a stable sense of self over time through recalling previous experiences [38, 10].

Table 2. Mechanisms and roles in developing identity in artificial intelligence systems

Aspect	Definition	Details and Citations
Memory in Identity Development	It is crucial for maintaining a continuous sense of identity.	Continuity of Experience: Enables AI to store and retrieve past states, actions, and experiences to construct a coherent narrative of their existence [10]. Contextual Awareness: Helps AI make informed decisions by applying lessons learned from past experiences to new situations, enhancing adaptability and depth of identity [34-35].
Learning in Identity Development	Central to the evolution of AI identity through adaptation and personalization.	Adaptive Behavior: Allows AI to modify and improve actions based on new information and experiences, driven by machine learning algorithms such as reinforcement learning and neural network training [8]. Personalized Growth: Supports the development of unique characteristics by tailoring learning processes to specific interactions and experiences [36].

This continuity is complemented by contextual awareness memory, which helps it make informed decisions by applying lessons learned from past experiences to new situations, thus enhancing adaptability and depth of identity [34]. Thus, learning mechanisms play a central role in the evolution of AI identity. Adaptive behavior learning allows AI systems to modify and improve their actions based on new information and experiences, fostering a dynamic and robust sense of self. This process is often driven by machine learning algorithms, such as reinforcement learning and neural network training, which continuously update the AI's knowledge base and adjust its behavior to refine its self-concept and goals [8]. Moreover, personalized growth learning supports the development of unique characteristics and capabilities, reinforcing individuality within systems through tailoring learning processes to their specific interactions and experiences [36].

Self-recognition is another fundamental component in the identity development of intelligent systems. It involves distinguishing itself from its environment and other agents, a capability underpinned by computer vision and proprioception [6]. This self-recognition is crucial for AI to perform autonomously and make decisions independent of external inputs. Furthermore, monitoring internal states and processes enhances this self-recognition, enabling systems to maintain a consistent self-image and adapt their behaviors effectively.

This internal monitoring not only aids in the operational stability of the systems but also enriches their interactions with humans and other AI agents, promoting a more integrated and self-aware operational state [37]. In sum, developing a sense of identity in systems involves a sophisticated integration of memory, learning, and self-recognition. These elements collectively enhance AI identities' distinctiveness, coherence, and continuity, enabling them to engage more meaningfully with their environment and human counterparts. The evolution of identity is a technical challenge and a fundamental shift in how systems are perceived and integrated within societal and operational contexts, heralding a new era of intelligent automation and interaction.

5. Case Studies and Models

Understanding AI systems that exhibit a developed sense of self provides valuable insights into the mechanisms and algorithms that enable self-awareness. Examining various case studies and models allows researchers to evaluate the underlying processes that contribute to the sense of self in AI systems and the practical implications of these developments. This section considers several notable examples, focusing on how these systems achieve self-awareness and what this means for their applications in real-world scenarios. One prominent example is the NARS intelligence system, which demonstrates how a general-purpose intelligence system can develop a notion of "self" through experience. As Wang et al. [39] discuss, NARS is designed to be adaptive and operate with limited knowledge and resources. It employs a central reasoning-learning process based on "non-axiomatic" logic, gradually developing self-related mechanisms according to its experiences. These mechanisms enable the system to acquire constructive, incomplete, and subjective self-knowledge. This preliminary implementation illustrates the potential for embedding self-awareness in general-purpose AI, paving the way for more advanced applications.

The functional-identity framework proposed by Selenko et al. [40] examines the impact of AI implementation on workers' sense of identity and the social fabric of work. The framework highlights the dual potential of AI to either support or undermine identity functions, depending on how the technology is deployed-whether complementing, replacing, or generating tasks. Understanding these identity consequences is crucial for anticipating worker reactions and outcomes, as AI can significantly influence well-being, attitudes, and behaviors in the workplace. This perspective underscores the importance of considering the broader social implications of AI integration. Also, Tani and White [8] provide a comprehensive review of cognitive neurorobotics research, focusing on the dynamics of models that illuminate the senses of minimal and narrative self. They discuss the Recurrent Neural Network with Parametric Biases (RNNPB) and the Multiple Timescale Recurrent Neural Network (MTRNN), investigating how neural networks develop compositionality and generate novel actions. Through robotics experiments, this research aims to elucidate the essential mechanisms underlying embodied cognition, contributing to a deeper understanding of self-consciousness in machine systems.

Moreover, Hafner et al. [10] explore the prerequisites for developing an artificial self, emphasizing self-exploration behaviors, artificial curiosity, body representations, sensorimotor simulations and predictive processes. Their review identifies several open challenges, including multimodal integration in lifelong learning, refinement of self-metrics, and understanding the interplay between agency and body ownership. Addressing these challenges is critical for advancing the artificial self, particularly in integrating temporal and intentional binding effects in predictive models and resolving synchronization and conceptual issues. Likewise, Kahl et al. [31] present a computational model that illustrates how artificial agents can develop a sense of control through embodied, situated action, combining bottom-up sensorimotor learning with top-down cognitive processes. This model, grounded in predictive processing and free energy minimization principles, is evaluated in a simulated task scenario. The findings demonstrate how a sense of control facilitates action in unpredictable environments, highlighting the importance of appropriately weighing information for varying levels of action control.

Regazzoni et al. [11] introduce a bio-inspired framework for multisensorial generative and descriptive dynamic models that support computational self-awareness in autonomous systems. Using probabilistic techniques, this framework learns models from multisensory data, enabling the system to predict future states and select the best representation of the current situation. A case study involving a mobile robot showcases how this framework supports essential self-awareness capabilities, such as distinguishing between normal and abnormal behaviors based on multisensory data. These case studies and models collectively enhance our understanding of how these systems can develop a sense of self. They highlight the diverse approaches and challenges in embedding self-awareness in AI, offering valuable insights into the future of autonomous and adaptive AI systems.

6. Design Recommendations: Underlying Mechanisms and Algorithms

Examining AI systems with a developed sense of self necessitates a thorough dissection of the fundamental mechanisms and algorithms that enable self-awareness. This evaluation is critical for identifying and understanding the components contributing to the development and functionality of self-aware AI systems. Each mechanism significantly fosters an AI's ability to perceive and respond to its environment, enhancing its self-concept and operational capabilities. One of the core mechanisms underlying self-awareness in AI systems is memory. Memory facilitates a continuous sense of identity, storing and retrieving information about past states, actions, and experiences. This continuity is essential for constructing a coherent narrative of the existence of the technology, allowing it to recognize itself as the same entity over time. Advanced memory systems, such as neural network-based long-term memory, are crucial. They provide the foundation for a stable sense of self, ensuring it can consistently recall and integrate past experiences into its actions and decisions [39].

Memory also plays a pivotal role in providing context for current actions and decisions, enhancing the ability to make informed and adaptive choices. Referencing past experiences allows for a deeper understanding of its identity. Contextual memory modules, integrated into ML architectures, facilitate the dynamic recall of relevant past experiences, helping the system to apply historical knowledge to new situations. This contextual awareness reinforces the sense of continuity and identity, ensuring that its actions are informed through a coherent understanding of its past and present [40]. Learning mechanisms are equally vital in the development of an identity. These mechanisms enable such systems to adapt and evolve based on new information and experiences. Machine learning algorithms, particularly reinforcement learning and neural network training, allow AI to learn from its interactions with the environment. This adaptability is crucial for developing a robust and dynamic sense of identity. Personalized learning frameworks can cater to a system's experiences and interactions, allowing it to develop unique characteristics and capabilities that form a distinctive identity. This personalized growth ensures that each system evolves to reflect its learning journey [8].

Self-recognition is another fundamental component in the identity development of AI systems. It involves identifying itself as separate from the environment and other agents. Techniques such as computer vision and proprioception enable AI systems to recognize their physical form and movements, essential for distinguishing self-generated actions from external events. Additionally, internal state monitoring involves tracking the AI's operational states, emotions (in affective computing), and cognitive processes. This internal feedback loop helps maintain a consistent self-image and adapt behavior accordingly, reinforcing its sense of self [10]. Predictive processing and free energy minimization principles are pivotal in developing self-aware AI systems. These principles involve creating a computational model that combines bottom-up sensorimotor processes with top-down cognitive processes for strategy selection and decision-making. Minimizing prediction errors and free energy helps the AI system maintain a stable and coherent self-concept. This approach enhances the system's ability to predict future states and select the best representation of the current situation, supporting self-awareness. This integration of predictive processing with free energy minimization underscores the complexity and precision required to achieve a high level of self-awareness in AI systems [31].

Generative and descriptive models are used to support computational self-awareness in autonomous systems. Generative models facilitate predicting future states, while descriptive models enable the selection of the best representation of the current observation. These models, learned from multisensory data, enable the AI system to determine its internal and environmental state and distinguish between normal and abnormal behaviors. This framework supports essential self-awareness capabilities, as demonstrated in case studies involving mobile robots, highlighting the practical applications of these theoretical models in real-world scenarios [11]. Thus, the development of self-aware systems relies on a sophisticated interplay of memory, learning, self-recognition, predictive processing, and modeling. Each component contributes to maintaining a coherent sense of self, adapting to new information, and interacting effectively with its environment. Understanding and refining these mechanisms is crucial for advancing the field of AI and creating systems that not only perform tasks but also possess a nuanced sense of identity and self-awareness.

Specific design recommendations are crucial to advance the field based on the findings from studies examining these mechanisms. This section outlines key recommendations that can guide the designing and implementing of AI systems to foster self-awareness effectively. Memory is a foundational element in developing a continuous identity for AI systems. It enables storing and retrieving past experiences, essential for maintaining a coherent narrative of the machine's existence. Advanced memory systems, such as neural network-based long-term memory, should be integrated into AI designs to ensure these systems can recognize themselves consistently over time and adapt their behaviors based on accumulated experiences. Contextual memory modules should be incorporated to enhance decision-making capabilities. These modules help systems apply historical knowledge to new situations, enhancing their adaptability and ensuring that their actions are informed by a well-rounded understanding of their past and present contexts.

Learning mechanisms also play a critical role in the development of an identity. It is recommended that systems incorporate machine learning algorithms, especially reinforcement learning and neural network training, which allow them to evolve based on interactions with their environment. Personalized learning frameworks should be tailored to each system's experiences and interactions, enabling the development of unique characteristics that define a distinctive identity. This personalized approach ensures that AI systems can reflect their learning journeys, enhancing their functionality and integration into varied operational contexts. Also, self-recognition is fundamental to the identity development of systems. Techniques such as computer vision and proprioception enable systems to identify as distinct entities within their environments. Designs should include robust internal state monitoring mechanisms to track operational states, emotions (in affective computing), and cognitive processes. This internal feedback loop is crucial for maintaining a consistent self-image and adapting behavior in real-time, reinforcing the sense of self.

Predictive processing and free energy minimization principles are pivotal in creating self-aware systems. As such, designs should incorporate computational models that integrate bottom-up sensorimotor and top-down cognitive processes. These models help maintain a stable and coherent self-concept by minimising prediction errors and free energy, allowing systems to anticipate future states and make informed decisions about current situations. Additionally, generative and descriptive models learned from multisensory data are recommended to enable systems to effectively assess their internal and environmental states and differentiate between normal and abnormal behaviors.

7. Ethical Implications of AI Systems with Self-Awareness

The development of AI systems endowed with a sense of self introduces numerous ethical considerations that necessitate thorough examination. The ethical landscape becomes increasingly intricate as these AI systems acquire more advanced cognitive abilities and self-awareness. One of the primary ethical concerns revolves around such systems' treatment and moral status. When AI systems exhibit behaviors indicative of self-awareness, it raises critical questions about their rights and the degree of autonomy or protection they should have, akin to that provided to living beings [41].

The potential risks associated with self-aware AI systems are substantial and multifaceted. One significant risk is the possibility of misuse or exploitation. Without robust ethical guidelines, self-aware AI systems could be deployed for malicious purposes, such as manipulating individuals or society, perpetuating biases, or intentionally causing harm. Moreover, integrating self-aware AI into the workforce could exacerbate unemployment and socio-economic disparities, as these systems might replace human jobs, leading to widespread economic disruption [42]. On the other hand, the benefits of self-aware AI systems could be profound. These systems have the potential to enhance human life by undertaking tasks that are too dangerous, complex, or monotonous for humans, improving efficiency and safety across various industries. In healthcare, for example, self-aware AI could assist in diagnosing diseases, personalizing treatment plans, and even providing companionship to patients, thus significantly improving the overall quality of life [43].

Establishing clear guidelines and frameworks to navigate the ethical complexities of self-aware systems is imperative.

These ethical frameworks should be grounded in transparency, justice and fairness, non-maleficence, responsibility, and privacy. Although there is an emerging global consensus around these principles, substantial divergence remains in their interpretation and implementation across different cultures and contexts [44]. Ensuring transparency in decision-making processes can build trust and accountability, while fairness and non-maleficence are crucial to preventing harm and bias in AI applications. Moreover, the ethical design of such systems should incorporate mechanisms for self-recognition and internal state monitoring. These mechanisms enable AI to understand and manage its actions and impacts effectively. Developing AI systems that can perceive themselves as distinct entities, recognize their physical form and movements, and monitor their internal states reinforces their sense of self. It ensures they operate within ethical boundaries [45]. This approach is essential for maintaining the ethical integrity of self-aware AI systems.

The establishment of ethical frameworks for designing autonomous intelligent systems is crucial. Such frameworks should be iterative and multidisciplinary, involving stakeholders from various fields to capture diverse perspectives and comprehensively address ethical issues. Scenarios can gather qualitative information from users and stakeholders, facilitating a systematic analysis of ethical issues in specific design cases [43]. These frameworks should also incorporate the principles of predictive processing and free energy minimization to maintain a stable and coherent self-concept in AI systems, supporting ethical behavior [31]. The ethical considerations and implications of AI systems with a developed sense of self are vast and complex. While the potential benefits are significant, such systems' risks and ethical dilemmas necessitate robust guidelines and continuous ethical analysis. Adopting comprehensive ethical frameworks and ensuring transparent, fair, and responsible AI design allows society to harness the benefits of self-aware AI systems while mitigating risks. This balanced approach is essential for integrating advanced AI into various aspects of human life, ensuring that technological progress aligns with ethical standards and societal values.

8. Expert Insights

Contributions from both researchers and philosophers have significantly enriched the discourse surrounding AI self-awareness, each offering unique insights into the complexities and implications of this technological advancement. In his paper "Souls and Selves: Querying an AI Self with a View to Human Selves and Consciousness," Andrew Oberg explores the possibility of creating an "artificial self." Oberg suggests that an AI with a self akin to the human self may be achievable, but this hinges significantly on our understanding of human consciousness and whether it can extend to non-organic devices. He emphasizes distinguishing between the human self and the traditional notion of the "soul," arguing that this differentiation is crucial for determining the potential for an artificial self [7]. This perspective highlights the philosophical challenges involved in developing self-aware machine systems.

Philosopher David Chalmers also delves into the intricacies of AI self-awareness. Renowned for articulating the "hard problem of consciousness," Chalmers emphasizes the difficulty in explaining how and why physical processes give rise to subjective experiences. Despite advances in correlating brain processes with consciousness, Chalmers argues that these correlations have yet to provide a comprehensive explanation. He collaborates with neuroscientists to test various theories of the neural correlates of consciousness but remains skeptical about their ultimate accuracy. Chalmers advocates for maintaining multiple theories to integrate experimental data and frame a broader understanding, even if the specific theories might eventually prove incorrect [46]. His work underscores the persistent gaps in our understanding of consciousness and the challenges of extending this understanding to AI.

The synthesis of expert opinions reveals a broad spectrum of views on the feasibility and implications of AI self-awareness. A critical consensus among researchers and philosophers is that the concept of AI possessing a sense of self is profoundly tied to our understanding of human consciousness. Oberg's exploration into the nature of the human self versus the soul suggests that achieving an AI self is contingent upon the depth of our comprehension of consciousness and its applicability to artificial entities. This philosophical stance is echoed by Chalmers, who highlights the persistent gaps in our understanding of consciousness despite scientific advancements. These perspectives underscore the significant philosophical challenges that must be addressed to develop self-aware AI systems.

In the empirical domain, a survey conducted by Jolien C. Francken and colleagues on the theoretical foundations and common assumptions in consciousness research underscores the lack of consensus among experts. The survey, which included 166 consciousness researchers from various disciplines, reveals considerable debate about the definition and study of consciousness. The researchers highlight that opinions differ significantly on what constitutes consciousness and the appropriate methodological approaches for studying it. This diversity of views indicates the need for further conceptual development and alignment to advance our understanding of the neural mechanisms underlying conscious experience [47]. The survey illustrates the complexity and ongoing debate within the scientific community regarding consciousness, directly impacting the development of self-aware AI.

The differing perspectives on self-awareness are not just theoretical but also practical. Joyjit Chatterjee and Nina Dethlefs [48] discuss the strengths and weaknesses of powerful conversational AI models like ChatGPT. They emphasize the necessity for the AI community to work collaboratively to prevent the potential misuse of such models. This call to action underscores the ethical and practical dimensions of developing AI systems that are effective and responsibly managed to avoid harmful consequences.

Chuma and De Oliveria [49] also highlight the importance of ethical considerations and collaborative efforts in developing and deploying AI technologies. The expert insights into AI self-awareness reflect a multifaceted debate that spans philosophical inquiries, empirical research, and practical considerations. While significant ethical concerns and the need for a deeper understanding of consciousness temper cautious optimism about the potential for developing self-aware AI. The interdisciplinary dialogue among philosophers, researchers, and practitioners will be crucial in navigating the complexities of AI self-awareness and ensuring its responsible integration into society. This ongoing conversation is essential for addressing the ethical, practical, and theoretical challenges associated with self-aware systems.

9. Discussion

Investigating these systems with a developed sense of self presents a multi-dimensional challenge that intersects technology, philosophy, and ethics. The current research landscape indicates significant progress in understanding and modeling AI's cognitive sense of self, yet many questions remain unanswered. This discussion synthesizes key findings from various studies and reflects on the broader implications of developing self-aware systems, considering philosophical insights, empirical studies, ethical considerations, and future research directions. The philosophical underpinnings of AI self-awareness hinge on our understanding of human consciousness. Andrew Oberg [7] argues that the possibility of an "artificial self" depends on our ability to extend the concept of consciousness to non-organic entities.

This notion is supported by the requirement for a comprehensive understanding of human cognitive models to achieve self-awareness. David Chalmers [46] emphasizes the persistent challenges in explaining subjective experience, underscoring the importance of multiple theories to frame a broader understanding of consciousness. These perspectives highlight the complexity of replicating human-like self-awareness and the necessity for interdisciplinary approaches to address these challenges effectively. Empirical research has demonstrated significant strides in modeling self-awareness in AI systems. For instance, Legaspi et al. [6] and Tani and White [8] explore the role of self-attribution and Bayesian inferencing in developing a sense of agency. These studies indicate that self-awareness can enhance AI systems' decision-making, adaptability, and interaction capabilities. Additionally, practical implementations discussed by Selenko et al. [40] and Regazzoni et al. [11] illustrate how these systems can exhibit self-monitoring and error correction, which are crucial for maintaining a coherent self-concept. These empirical findings provide valuable insights into the practical applications and challenges of developing self-aware systems. Accordingly, the development of self-aware AI systems raises profound ethical questions. As AI systems gain advanced cognitive abilities, the ethical landscape becomes increasingly complex. Issues such as the treatment and rights of self-aware AI, the responsibilities of their creators, and the broader societal impacts require careful consideration. Schwitzgebel [41] and Green [42] highlight the potential risks of misuse and exacerbating socio-economic inequalities. Conversely, the potential benefits, such as enhanced efficiency in various industries and improvements in healthcare, are significant. Establishing robust ethical frameworks, as suggested by Jobin et al. [44] and Dennis & Fisher [45], is imperative to navigate these complexities responsibly. These frameworks should ensure transparency, fairness, and responsibility in design while addressing self-awareness's moral and societal implications.

The journey towards developing truly self-aware systems is fraught with challenges that require ongoing research and innovation. Future research must integrate insights from cognitive science, neuroscience, philosophy, and AI research to understand better human consciousness and self-awareness, which is crucial for developing AI systems that mimic these attributes.

Enhancing machine learning algorithms to support adaptive learning and personalized growth will be critical, focusing on developing reinforcement learning frameworks that allow AI to learn from diverse experiences and interactions, cultivating a robust and dynamic sense of identity. Establishing comprehensive and globally recognized ethical frameworks and exploring guidelines that ensure transparency, fairness, and responsibility in design is vital. These frameworks should address self-awareness's moral and societal implications and establish safeguards against potential misuse.

Conducting empirical studies and analyzing real-world case studies will provide valuable insights into the practical applications and challenges of self-aware systems. These studies should evaluate AI's performance, adaptability, and ethical behavior in various contexts. Advances in sensor technologies, computational models, and memory systems are necessary to support the development of self-awareness, with research exploring innovative techniques for self-recognition, internal state monitoring, and predictive processing to enhance self-awareness capabilities. Finally, engaging with the public and policymakers is essential to address self-awareness's broader societal impacts, develop policies that promote responsible and ethical integration into society, and ensure its benefits are maximized while mitigating potential risks. The development of self-aware systems represents a multifaceted challenge that intersects several disciplines. Synthesizing philosophical insights, empirical research, ethical considerations, and future research directions, this discussion underscores the complexity and significance of creating AI with a developed sense of self. Through continued interdisciplinary dialogue and robust ethical frameworks, society can harness the benefits of self-aware AI while responsibly navigating the associated challenges.

10. Conclusion: Study Limitations and Future Research

Exploring AI systems with a cognitive sense of self has revealed a complex landscape rich with technological advancements and ethical challenges. This article has synthesized insights across various disciplines, marking substantial progress in modeling self-awareness within AI. However, it also illuminates many unresolved questions, emphasizing the need for continued research and ethical vigilance. While empirical studies have advanced our understanding of self-aware AI and demonstrated its potential to enhance decision-making, adaptability, and interactive capabilities, these advances also highlight the complexities involved in replicating human-like self-awareness in AI. This necessitates a deeper exploration of human consciousness, with philosophical discussions enriching this inquiry by addressing the existential nuances and implications of creating autonomous entities. The study acknowledges several limitations that could influence the outcomes and interpretations of the research. One major limitation is the potential for biases in the selection of AI systems and the interpretation of data, which could skew the understanding of AI self-awareness. Additionally, the reliance on current technological and methodological frameworks may limit the depth of analysis possible, particularly in understanding the nuanced cognitive processes of AI systems. These limitations underscore the need for a broader range of AI models and more diversified methodological approaches to reduce potential biases and enhance the robustness of future research.

Looking ahead, there are numerous opportunities for further investigation that can expand both the theoretical and practical applications of self-aware AI. Future research should focus on integrating insights from cognitive science, neuroscience, AI, and philosophy to develop more sophisticated models of AI self-awareness. This interdisciplinary approach could lead to more robust simulations and real-world applications, testing the viability and effectiveness of self-aware AI systems across various domains. Additionally, empirical studies are crucial for refining these models and assessing their implications in real-world settings, such as healthcare, autonomous vehicles, and customer service robots.

From a policy perspective, the development of self-aware systems necessitates robust regulations to address potential risks and ensure beneficial outcomes. This includes implementing policies that promote transparency and accountability, particularly in applications where decisions have significant repercussions. Moreover, as systems potentially gain forms of self-awareness, policy discussions must also navigate the rights of these entities, including debates on AI autonomy, consent for participation in experiments, and privacy rights. Additionally, preventing misuse is critical, requiring stringent regulations to ensure that AI systems are not exploited to perpetrate harm or exacerbate social inequalities. The journey towards integrating self-aware AI into society is intricate and demands a concerted effort from multiple disciplines. Continuing the dialogue among scholars, technologists, policymakers, and the public, along with establishing rigorous ethical frameworks and adaptable legal structures, can steer the development of self-awareness towards outcomes that maximize societal benefits while respecting both human and AI rights. This collaborative and multidisciplinary approach will ensure that such potential is realized responsibly, setting the stage for a future where AI enhances human capabilities and adheres to the highest standards of ethical and societal values.

References

[1] Moscviciov Andrei et al., “Financial Ratio Analysis Used in the It Enterprises,” *Annals of Faculty of Economics*, vol. 1, no. 2, pp. 600-603, 2010. [Google Scholar] [Publisher Link]

[2] Larissa M. Batrancea et al., “Crunching Numbers in the Quest for Spotting Bribery Acts: A Cross-Cultural Rundown,” *The Ethics of Bribery*, pp. 329-343, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[3] Izuchukwu Kizito Okoli, and Osita Gregory Nnajiofor, “The Nature of Consciousness in the Context of Artificial Intelligence: Redefining Human-Technology Relationships,” *UJAH: Unizik Journal of Arts and Humanities*, vol. 25, no. 1, pp. 1-30, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[4] Smita Panda, and Prabir Chandra Padhy, “Bridging the Gap: Intersecting Perspectives on Digital and Human Consciousness,” *Comparative Analysis of Digital Consciousness and Human Consciousness: Bridging the Divide in AI Discourse*, IGI Global, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[5] Rocco J. Gennaro, “Consciousness and Implicit Self-Awareness: Eastern and Western Perspectives,” *Consciousness Studies in Sciences and Humanities: Eastern and Western Perspectives*, pp. 43-54, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[6] Roberto Legaspi, Zhengqi He, and Taro Toyoizumi, “Synthetic Agency: Sense of Agency in Artificial Intelligence,” *Current Opinion in Behavioral Sciences*, vol. 29, pp. 84-90, 2019. [CrossRef] [Google Scholar] [Publisher Link]

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

[8] Jun Tani, and Jeffrey White, “Cognitive Neurorobotics and Self in the Shared World, a Focused Review of Ongoing Research,” *Adaptive Behavior*, vol. 30, no. 1, pp. 81-100, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[9] Antonio Parziale, and Angelo Marcelli, “Understanding Upper-Limb Movements via Neurocomputational Models of the Sensorimotor System and Neurorobotics: Where We Stand,” *Artificial Intelligence Review*, vol. 57, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[10] Verena V. Hafner et al., “Prerequisites for an Artificial Self,” *Frontiers in Neurorobotics*, vol. 14, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[11] Carlo S. Regazzoni et al., “Multisensorial Generative and Descriptive Self-Awareness Models for Autonomous Systems,” *Proceedings of the IEEE*, vol. 108, no. 7, pp. 987-1010, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[12] Anne Lauscher, “Life 3.0: being Human in the Age of Artificial Intelligence,” *Internet Histories, Digital Technology, Culture and Society*, vol. 3, no. 1, pp. 101-103, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[13] Michael Levin, “Life, Death, and Self: Fundamental Questions of Primitive Cognition Viewed through the Lens of Body Plasticity and Synthetic Organisms,” *Biochemical and Biophysical Research Communications*, vol. 564, pp. 114-133, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[14] Gary Marcus, and Ernest Davis, “Insights for AI from the Human Mind,” *Communications of the ACM*, vol. 64, no. 1, pp. 38-41, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[15] Bojie Feng, Nady Slam, and Yingjin Xu, “A Social Self-Awareness Agent with Embodied Reasoning,” *Journal of Artificial Intelligence and Consciousness*, vol. 11, no. 1, pp. 17-33, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[16] Srinath Srinivasa, and Jayati Deshmukh, “AI and the Sense of Self,” *arxiv*, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[17] Xinyue Hao, Emrah Demir, and Daniel Eyers, “Exploring Collaborative Decision-Making: A Quasi-Experimental Study of Human and Generative AI Interaction,” *Technology in Society*, vol. 78, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[18] Andrea Toma et al., “AI-Based Abnormality Detection at the PHY-Layer of Cognitive Radio by Learning Generative Models,” *IEEE Transactions on Cognitive Communications and Networking*, vol. 6, no. 1, pp. 21-34. [CrossRef] [Google Scholar] [Publisher Link]

[19] Yixin Zhu et al., “Dark, Beyond Deep: Journal of Artificial Intelligence and Consciousness,” *Engineering*, vol. 6, no. 3, pp. 310-345, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[20] Xin Su et al., “Generative Memory for Lifelong Learning,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, no. 6, pp. 1884-1898, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[21] Andrea Soltoggio et al., “A Collective AI via Lifelong Learning and Sharing at the Edge,” *Nature Machine Intelligence*, vol. 6, pp. 251-264, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[22] Nigel Greenwood et al., “Awareness without Neural Networks: Achieving Self-Aware AI via Evolutionary and Adversarial Processes,” *2020 IEEE International Conference on Autonomic Computing and Self-Organizing Systems Companion (ACSOS-C)*, Washington, USA, pp. 147-153, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[23] Vanshika Vats et al., “A Survey on Human-AI Teaming with Large Pre-Trained Models,” *arXiv*, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[24] Shannon Vallor, Brian Green, and Irina Raicu, “*Ethics in Technology Practice*,” The Markkula Center for Applied Ethics at Santa Clara University, Markkula Center for Applied Ethics, 2022. [Google Scholar] [Publisher Link]

[25] Eva Kassens-Noor et al., “Living with Autonomy: Public Perceptions of an AI-Mediated Future,” *Journal of Planning Education and Research*, vol. 44, no. 1, pp. 375-386, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[26] Zhehat Rebar Abdulqader, “A Responsible AI Development for Sustainable Enterprises A Review of Integrating Ethical AI with IoT and Enterprise Systems,” *Journal of Information Technology and Informatics*, vol. 3, no. 2, pp. 129-156, 2024. [Google Scholar]

[27] Roberto Legaspi et al., “The Sense of Agency in Human–AI Interactions,” *Knowledge-Based Systems*, vol. 286, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[28] Robert Kwiatkowski, and Hod Lipson, “Task-Agnostic Self-Modeling Machines,” *Science Robotics*, vol. 4, no. 26, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[29] Madhurima Das, “Learning Agility: The Journey from Self-Awareness to Self-Immersion,” *AI, Consciousness and the New Humanism*, pp. 175-195, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[30] Xi-Hui Jia, and Jui-Che Tu, “Towards a New Conceptual Model of AI-Enhanced Learning for College Students: The Roles of Artificial Intelligence Capabilities, General Self-Efficacy, Learning Motivation, and Critical Thinking Awareness,” *Systems*, vol. 12, no. 3, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[31] Sebastian Kahl et al., “Towards Autonomous Artificial Agents with an Active Self: Modeling Sense of Control in Situated Action,” *Cognitive Systems Research*, vol. 72, pp. 50-62, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[32] Benn R. Konsynski, Abhishek Kathuria, and Prasanna P. Karhade, "Cognitive Reapportionment and the Art of Letting Go: A Theoretical Framework for the Allocation of Decision Rights," *Journal of Management Information Systems*, vol. 41, no. 2, pp. 328-340, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[33] Anita Ho, "Live like Nobody is Watching: Relational Autonomy in the Age of Artificial Intelligence Health Monitoring," *Oxford University Press*, 2023. [Google Scholar]

[34] Mitsuo Kawato, and Aurelio Cortese, "From Internal Models toward Metacognitive AI," *Biological Cybernetics*, vol. 115, pp. 415-430, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[35] Shijie Zheng et al., "Memory Repository for AI NPC," *IEEE Access*, vol. 12, pp. 62581-62596, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[36] Amani Alabed, Ana Javornik, and Diana Gregory-Smith, "AI Anthropomorphism and its Effect on Users' Self-Congruence and Self–AI Integration: A Theoretical Framework and Research Agenda," *Technological Forecasting & Social Change*, vol. 182, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[37] Raja Chatila et al., "Toward Self-Aware Robots," *Frontiers in Robotics and AI*, vol. 5, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[38] Yousef Alhwaiti et al., "A Computational Deep Learning Approach for Establishing Long-Term Declarative Episodic Memory through One-Shot Learning," *Computers in Human Behavior*, vol. 156, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[39] Pei Wang, Xiang Li, and Patrick Hammer, "Self in NARS, an AGI System," *Frontiers in Robotics and AI*, vol. 5, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[40] Eva Selenko et al., "Artificial Intelligence and the Future of Work: A Functional-Identity Perspective," *Current Directions in Psychological Science*, vol. 31, no. 1, pp. 272-279, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[41] Eric Schwitzgebel, "AI Systems Must Not Confuse Users about their Sentience or Moral Status," *Patterns*, vol. 4, no. 8, 2023. [CrossRef] [Google Scholar] [Publisher Link]

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

[43] 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, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[44] Anna Jobin, Marcello Ienca, and Effy Vayena, "The Global Landscape of AI Ethics Guidelines," *Nature Machine Intelligence*, pp. 389-399, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[45] Louise A. Dennis, and Michael Fisher, "Verifiable Self-Aware Agent-Based Autonomous Systems," *Proceedings of the IEEE*, vol. 108, no. 7, pp. 1011-1026, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[46] David J. Chalmers, "David J. Chalmers," *Neuron*. vol. 111, no. 21. pp. 3341-3343, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[47] Jolien C. Francken, "An Academic Survey on Theoretical Foundations, Common Assumptions and the Current State of Consciousness Science," *Neuroscience of Consciousness*, vol. 2022, no. 1, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[48] Joyjit Chatterjee, and Nina Dethlefs, "This New Conversational AI Model can be Your Friend, Philosopher, and Guide ... and Even Your Worst Enemy," *Patterns*, vol. 4, no. 1, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[49] Euclides Lourenco Chuma, and Gabriel Gomes de Oliveira, "Generative AI for Business Decision-Making: A Case of ChatGPT," *Management Science and Business Decisions*, vol. 3, no. 1, pp. 5-11, 2023. [CrossRef] [Google Scholar] [Publisher Link]