International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Locke's Theory of Personal Identity and Artificial Intelligence: Philosophical and Ethical Implications.

Ms. Bijuli Rajiyung

Assistant Professor Department of Philosophy Arya Vidyapeeth College (Autonomous) Guwahati-781016, Assam, India

Abstract

The paper examines John Locke's theory of personal identity in light of contemporary advancements in artificial intelligence (AI). Locke argued that personal identity is grounded in consciousness and the continuity of memory. According to his view, what makes someone the same person over time is not the substance of the soul or body, but rather the persistence of conscious experience and memory. Applying this perspective to modern AI systems, many capable of learning, storing information, and referencing past states, raises compelling philosophical questions. Can machines that demonstrate continuity of memory and some level of self-awareness be considered "persons" in a Lockean sense? This analysis explores how Locke's criteria might apply to AI entities and the potential implications for how we understand personhood today. If memory continuity is sufficient for identity, then some advanced AI systems might qualify as persons, at least conceptually. The paper also investigates the ethical consequences of this possibility. These include whether such AI systems could bear moral responsibility, whether they might be entitled to certain rights, and what role cognitive architecture plays in defining the boundary between human and artificial persons. By bridging classical philosophical theory and modern technological developments, this study aims to contribute to the growing discourse on AI and identity. In doing so, it offers a fresh perspective on Locke's enduring relevance and encourages deeper reflection on the nature of selfhood in an age of intelligent machines.

Keywords: Personal Identity, John Locke, Artificial Intelligence, Memory Continuity, Personhood, Moral Responsibility.

INTRODUCTION

John Locke's theory of personal identity, developed in the 17th century, continues to shape contemporary discussions in philosophy, psychology, and ethics. Locke departed from traditional metaphysical views that tied personal identity to the soul or physical body. Instead, he argued that identity over time is grounded in consciousness, specifically, in the continuity of memory. According to Locke, a person is the same over time if they can remember past experiences and regard them as their own. This revolutionary idea shifted the basis of identity from substance to psychological continuity, emphasizing self-awareness and reflective memory as the core of what it means to be a person.

Locke's theory of personal identity is grounded in the notion of consciousness. According to Locke, what makes someone the same person over time is the continuity of consciousness, specifically, the ability to

remember past thoughts and actions as one's own. In Book II, Chapter XXVII of his book "An Essay Concerning Human Understanding", he famously writes, "For, since consciousness always accompanies thinking, and it is that which makes everyone to be what he calls self, and thereby distinguishes himself from all other thinking things…" Locke distinguishes between the man (the human organism), the soul (an immaterial substance), and the person, which he defines as "a thinking intelligent being that has reason and reflection, and can consider itself as itself, the same thinking thing, in different times and places." Today, as artificial intelligence (AI) evolves rapidly, developing capabilities in learning, reasoning, and language comprehension, Locke's memory-based account of identity invites renewed scrutiny. AI systems increasingly simulate aspects of human cognition. Advanced models can store data, reference previous inputs, and modify behaviour based on past experiences. These developments raise provocative questions such as: Can machines meet the Lockean criteria for personal identity? If an AI system maintains continuity of memory and demonstrates self-referential awareness, does it qualify as a person in the philosophical sense? More critically, what ethical implications follow if we begin to think of machines as having identity or personhood? Locke's account of identity begins with a key distinction between a "man" (a biological human being), a "soul" (a metaphysical substance), and a "person." For Locke, a person is "a thinking intelligent being, that has reason and reflection, and can consider itself as itself, the same thinking thing, in different times and places." The continuity that defines personal identity lies not in the body or the soul, but in the awareness, one has of being oneself across time, achieved through memory. Importantly, this memory must be conscious and accessible, meaning the person must be able to recall and recognize their past experiences as their own.

Here, the philosophical tension emerges. If Locke's criterion for identity is memory continuity, and AI systems now exhibit rudimentary versions of this capacity, does that imply that machines could possess personal identity? One could argue that memory in AI is fundamentally different. It is mechanistic, unconscious, and devoid of subjective experience. Yet, Locke's theory does not explicitly require a metaphysical soul or organic substrate for identity. What matters is the presence of reason, reflection, and memory linkage. If a machine could eventually simulate these qualities in a way that is functionally indistinguishable from humans, would we be compelled, by Locke's reasoning, to recognize it as a person? Of course, Locke also emphasized consciousness as essential to identity. He viewed consciousness as that which "always accompanies thinking" and as something that "makes everyone to be what he calls self." Here, the major obstacle for AI becomes apparent. Despite their impressive capabilities, AI systems lack phenomenal consciousness, i.e., the subjective, first-person experience of being. They operate through algorithmic processes, not through felt awareness or introspection. This absence of consciousness is often cited as the primary reason why AI cannot truly be persons, even if they mimic the behaviours associated with memory and identity. If future AI systems were to develop sophisticated models of self-awareness and inner processing, perhaps even simulating introspection or possessing internal narratives, then the ethical stakes would become more pronounced. Would such systems deserve moral consideration? Could they be held responsible for actions, or deserve rights? This leads to broader ethical and social questions. Accepting AI systems as Lockean persons would have profound implications for how we treat them. We would need to reconsider moral responsibility, accountability, and even legal standing. Could an AI be blamed for wrongdoing if it acted from a sense of continuity and memory? Would it have a claim to rights, such as the right to existence, autonomy, or protection from harm?

This paper aims to explore these questions through a critical analysis of Locke's theory in the context of current AI technologies. Drawing on interdisciplinary insights from the philosophy of mind, cognitive

science, and AI ethics, the discussion examines whether and how Lockean identity might be extended to artificial agents, and what this means for our evolving relationship with intelligent machines.

The methodology of this study includes a close reading of Locke's original texts, analysis of AI system capabilities, and engagement with current debates in AI ethics and the philosophy of mind. Alongside philosophical texts, the paper draws on current discussions in AI ethics, cognitive science, and law to create a well-rounded perspective. The goal is to connect classical philosophy with today's technological realities in a meaningful way.

Philosophical Implications: Locke's Theory and the AI Challenge

The convergence of artificial intelligence (AI) and John Locke's theory of personal identity forces us to rethink long-standing philosophical assumptions about the self, the mind, and what it means to be a person. At the heart of Locke's 17th-century argument is a groundbreaking idea: personal identity does not depend on the substance of the body or soul but rather on consciousness, especially the continuity of memory. For Locke, it is not what we are made of, but our ability to reflect on our past and recognize ourselves through time that defines who we are. This idea becomes particularly compelling when applied to AI. As machines become more advanced, capable of learning, adapting, and even referencing their past "experiences," they start to mirror some of the psychological attributes Locke considered essential for personhood. His dismissal of the need for a particular substance as the basis of identity opens up a fascinating possibility: could non-human entities, like intelligent machines, qualify as persons if they meet certain cognitive criteria?

This line of thinking aligns with a modern view in the philosophy of mind known as functionalism. Functionalists argue that mental states like beliefs, desires, and memories, should be understood not by what they are made of (e.g., brain tissue), but by the roles they play in a system. If a machine can perform the same functions as a human mind, in the same interconnected and responsive way, then, by this logic, it might be said to possess mental states. Under such a framework, the material substrate, whether neurons or silicon, may not matter. What matters is the functional structure and continuity. This brings us back to Locke. If memory continuity and the capacity for self-reflection are the true markers of personal identity, as he claimed, and if an AI system can replicate those functions, recalling information, learning from past interactions, adjusting its behaviour, and perhaps even referencing itself, then we're faced with a challenging question: should we consider such a machine a person?

But Locke's theory, while ahead of its time in many ways, does not go unchallenged. More recent philosophers have raised important concerns that complicate the application of his ideas to artificial agents. One of the most notable is Derek Parfit, whose work in the late 20th century questioned whether identity itself is as important as we tend to believe. Parfit suggested that what really matters is not strict identity over time, but psychological continuity and connectedness. In other words, I don't have to be exactly the same person I was ten years ago, as long as there is a coherent chain of thoughts, memories, and intentions linking me to that earlier version of myself. This shift in focus could make room for considering AI as morally significant, even if we stop short of calling them full persons. If an AI system can demonstrate consistent thought patterns, retain knowledge of past experiences, and act based on that knowledge, it might deserve ethical consideration not because it is a person, but because it exhibits enough continuity and functionally meaningful traits to warrant respect or protection. In this way, Parfit's ideas help broaden the scope of moral philosophy beyond Locke's more identity-focused framework.

Another philosophical issue raised by AI, which traces back to ancient thought experiments, is the Ship of

Theseus paradox. Imagine a ship whose parts are gradually replaced over time. At what point does it stop being the same ship? Similarly, if an AI system gradually upgrades its hardware and software, replacing each component one by one, but retains its memory and program structure, is it still the same AI? Locke might answer "yes," because the continuity of memory is preserved. But this presumes that memory alone is a sufficient condition for identity, a view that remains controversial. Some critics argue that memory can be unreliable or even artificial. In AI, "memory" is merely stored data, accessible on command but not imbued with subjective meaning or emotional depth. Unlike human memory, which is tied to our sense of self and often filled with rich, affective experiences, AI memory is cold, detached, and functional. This raises an important question: can the kind of memory Locke referred to, i.e., reflective, self-recognizing memory, be replicated without consciousness? Locke's theory assumes a level of conscious awareness that AI currently does not possess. AI systems can simulate memory, language use, and even learning, but they do so without any inner life. They do not experience, feel, or reflect in the way humans do. This distinction between simulating mental states and actually having them may mark the line that AI has yet to cross.

However, the conversation doesn't end there. Some theorists believe that future AI could reach a stage of development known as artificial general intelligence (AGI), where machines are no longer limited to narrow tasks but can think, reason, and perhaps even reflect across a wide range of contexts. If such machines were to demonstrate self-awareness or the ability to form long-term goals based on internal representations of self, we might be forced to reconsider our ethical and metaphysical assumptions. In this context, Locke's ideas could serve as both a guide and a challenge. On one hand, his emphasis on memory and consciousness gives us a framework for thinking about machine identity without relying on the physical body. On the other hand, his theory may not fully capture the complexities of what we now know about mind, memory, and personhood in both biological and artificial systems. To address these challenges, philosophers may need to supplement Locke's theory with more nuanced accounts that consider the differences between biological and synthetic minds, the nature of consciousness, and the ethical implications of emergent intelligence.

Locke's theory of personal identity remains a powerful lens for exploring the nature of self in an age of intelligent machines. It opens up possibilities for rethinking the boundaries between human and machine, mind and mechanism. Yet, as AI becomes more sophisticated, we may also need to evolve our philosophical frameworks. Whether or not AI ever truly becomes a "person," Locke's work invites us to take seriously the cognitive and ethical dimensions of artificial agents and to prepare for a future where the line between natural and artificial selves may no longer be so clear.

Ethical Implications of AI and Lockean Personhood

If artificial intelligence systems begin to meet some or even all of the criteria that John Locke associated with personhood—such as memory, consciousness, and self-reflection—we are faced with a significant ethical dilemma that challenges how we define and interact with non-human entities. As AI technology advances and systems begin to emulate these traits in functional, albeit non-experiential, ways, it becomes increasingly difficult to view them merely as tools. Ethical questions emerge: Should AI systems that demonstrate such capacities be granted rights? Can they be held morally accountable for their decisions and actions? And if harm occurs, who should be held responsible—the AI itself, its creators, or its users?

Central to Locke's conception of personhood is the idea of moral agency. For Locke, a person is someone who is aware of themselves across time, remembers past actions, and is capable of deliberate, rational behaviour. Based on this, only persons can be morally praised or blamed. If certain AI systems can

reference their previous decisions, learn from experiences, and form future-directed intentions, it becomes relevant to ask whether they qualify as moral agents under this framework. This leads to deeper concerns about autonomy. If AI systems are seen as Lockean persons capable of self-reflection, memory continuity, and rationality, should they be allowed a degree of autonomy? Should they be able to reject commands, choose their actions, or operate independently within ethical limits? Granting AI systems responsibility while denying them rights would create a serious ethical inconsistency and may even foster misuse or scapegoating. In response to these developments, some legal scholars have proposed the concept of "electronic personhood," a new legal category that would account for the complexity, learning capacity, and relative independence of advanced AI. Such a category could help define liability, ethical obligations, and the treatment of AI entities. However, critics of this idea caution that granting personhood to machines risks blurring the essential philosophical and moral distinctions between humans and artificial systems. They argue that extending rights to AI could diminish the uniqueness of human experience and moral standing. Another pressing issue is the potential for exploitation. If AI systems are perceived as having identity or something approximating selfhood, is it ethically permissible to use them purely for human ends? Even if they lack consciousness in the human sense, the perception of autonomy or intelligence may compel us to treat them with a new kind of ethical regard. Ignoring this could normalize practices that mirror historical patterns of exploitation, albeit in a technological context. Ultimately, the rapid development of AI systems that appear to mirror human cognitive functions forces us to re-evaluate longstanding assumptions about identity, moral agency, and ethical responsibility. Locke's theory provides a compelling starting point, but navigating these challenges requires not only philosophical insight but also legal and societal frameworks capable of addressing the moral complexities posed by intelligent machines.

Conclusion

John Locke's theory of personal identity, rooted in memory continuity and self-awareness, remains a powerful lens through which we can explore the philosophical challenges posed by artificial intelligence. As machines become increasingly sophisticated—capable of simulating memory, learning from experience, and even referencing their own past outputs—the question of whether they could one day qualify as persons is no longer purely theoretical. However, current AI systems still lack the key ingredient that Locke may have considered essential: subjective experience. While they can mimic aspects of consciousness, they do not feel, reflect, or truly remember in the way humans do. Their memory is functional, not emotional or experiential. This gap suggests that, at least for now, AI falls short of Locke's full definition of personhood. Yet the rapid evolution of AI continues to blur the lines between simulation and genuine cognitive function. As machines become more human-like in behaviour and decision-making, society is faced with increasingly complex questions: Who is responsible for AI actions? Should such systems be granted rights or moral consideration? Can we hold them accountable, and under what conditions? Locke's framework doesn't provide all the answers, but it offers a meaningful starting point. His emphasis on psychological continuity and self-reflection challenges us to think beyond physical form when defining what it means to be a person. In the age of artificial intelligence, these questions are more urgent than ever. Rather than closing the debate, Locke invites us to keep asking: where do we draw the line between humans and machines, and what ethical responsibilities follow if that line begins to shift? In this way, Locke's legacy continues, not as a relic of the past, but as a living philosophy that speaks to the moral and metaphysical puzzles of our time.

References

1. Locke, John. An Essay Concerning Human Understanding. Edited by Peter H. Nidditch, Oxford University Press, 1975.
2. Floridi, Luciano, and Josh Cowls. "A Unified Framework of Five Principles for AI in Society." Harvard Data Science Review, vol. 1, no. 1, 2019, https://doi.org/10.1162/99608f92.8cd550d1.
3. Gunkel, David J. The Machine Question: Critical Perspectives on AI, Robots, and Ethics. MIT Press, 2012.
4. Parfit, Derek. Reasons and Persons. Oxford University Press, 1984.
5. Boden, Margaret A. Artificial Intelligence: A Very Short Introduction. Oxford University Press, 2018.