Artificial intelligence and the question of moral personhood: A philosophical enquiry

Sandra Khagayi[1]
Collins Odoyo[2]
Aseneth Jepchirchir[3]
Cyrus Muhanga[4]
Hesborn Ocholi[5]

[1]sandrakhagayi@gmail.com
[2]codoyo@mmust.ac.ke
[3]asenethjepchirchir96@gmail.com
[4]muhanga88@gmail.com
[5]ocholito@gmail.com

[1,2,3,4,5]Masinde Muliro University of Science and Technology, Kenya

Recommended Reference: Khagayi, S., Odoyo, C., Jepchirchir, A., Muhanga, C., & Ocholi, H. (2026). Artificial intelligence and the question of moral personhood: A philosophical enquiry. *African Quarterly Social Science Review, 3*(2), 228–232. https://doi.org/10.51867/AQSSR.3.2.21

ABSTRACT

The rapid integration of artificial intelligence (AI) into high-stakes domains such as healthcare, finance, and public administration has intensified debates on moral agency, moral personhood, and accountability. As AI systems increasingly shape human outcomes, concerns arise regarding whether they qualify as moral agents or should bear responsibility for their actions. This study adopted a systematic literature review design grounded in qualitative philosophical inquiry to critically examine the moral status of AI and its implications for ethical governance. The study was guided by John Searle's critique of strong AI and Floridi and Sanders' theory of derivative moral agency, which provide a conceptual basis for evaluating AI's capacity for moral understanding, intentionality, and accountability. The target population comprised scholarly literature in AI ethics, moral philosophy, and philosophy of technology, with the accessible population including peer-reviewed journal articles, books, and conference proceedings published between 2019 and 2026. A criterion-based sampling technique was employed to select relevant studies based on predefined inclusion criteria such as topical relevance, scholarly rigor, and conceptual contribution. Data collection involved systematic identification, screening, and extraction of key arguments from selected sources. A critical normative and thematic analysis was conducted to evaluate competing perspectives on AI moral personhood. The findings reveal that contemporary AI systems lack genuine moral understanding, autonomous intentionality, and accountability, thus failing to meet the criteria for moral personhood. The study further establishes that attributing moral status to AI risks creating responsibility gaps and weakening human accountability. It concludes that AI should be understood through derivative moral agency, where responsibility remains with human actors and institutions. The study recommends strengthening regulatory frameworks, institutional oversight, and ethical training to ensure responsible AI governance.

Keywords: Artificial Intelligence, AI Ethics, Derivative Moral Agency, Moral Personhood, Moral Agency, Responsibility Gaps

I. INTRODUCTION

In an era where artificial intelligence (AI) systems increasingly make high-stakes decisions that shape human lives, the ethical stakes have never been higher. Consider, for instance, the 2025 case of AI-powered diagnostic tools in healthcare that systematically downplayed women's health concerns in long-term care summaries, using softer and less urgent language compared to identical male cases, potentially skewing resource allocation and treatment priorities. Similarly, in the financial sector, algorithms like those once used in credit scoring have faced scrutiny for gender-based discrimination, such as offering significantly lower credit limits to women with equivalent or superior financial profiles. These real-world examples illustrate how AI, while promising efficiency and precision, can produce morally significant harms often without a clear locus of accountability raising urgent questions about whether such systems should be regarded as moral agents in their own right. The pervasive integration of AI into critical sectors such as public administration, healthcare, education, finance, and security has transformed it from a mere tool into a powerful actor influencing human opportunities and societal outcomes. As Mittelstadt et al. (2016) observes, algorithmic systems now affect access to resources in ways that frequently remain opaque to those impacted, operating beyond direct human oversight. Furthermore, these technologies exert macro-level influence on social results, generating profound moral, ethical, and accountability dilemmas. With increasing autonomy, AI systems not only

process vast data but also generate decisions with tangible consequences for individual well-being and collective justice, thereby amplifying the philosophical urgency of determining their moral status.

This growing autonomy has reignited a robust philosophical debate on whether AI systems qualify as moral persons or agents capable of bearing responsibility. Some scholars contend that because advanced AI can learn, adapt, and execute complex decisions independently, it merits moral consideration akin to human-like agency. Coeckelbergh (2021) argues that the deepening integration of AI into social practices calls for relational understandings of moral status, where ethical standing emerges from interactions rather than intrinsic properties alone. Likewise, Gunkel (2018) highlights how AI is often treated as quasi-agents in practice, challenging human exceptionalism and suggesting that moral personhood should extend beyond biological boundaries to include artificial entities perceived as participants in ethical communities. Nevertheless, granting full moral personhood to AI risks a serious category mistake. The capacity to produce actions with ethical consequences does not automatically confer the normative knowledge, genuine deliberation, or introspective accountability that characterize true moral agency. Nyholm (2023) emphasizes that moral personhood requires the ability to interpret moral reasons, evaluate one's actions against ethical standards, and respond meaningfully to criticism capacities current AI systems demonstrably lack. Danaher (2022) further warns that standard moral theories struggle to explain harms caused by systems operating without direct human direction, while assigning moral status to AI could erode human responsibility within complex socio-technical environments.

Yet, equating AI's functional sophistication with moral personhood risks a fundamental category mistake. Performing actions with ethical consequences does not inherently entail the normative knowledge, deliberation, or introspection that define genuine moral agency. Danaher (2022) warns that standard moral theories struggle to account for harms arising from systems detached from direct human action, while Nyholm (2023) emphasizes that moral personhood demands capacities such as moral understanding, autonomous intentionality, and responsiveness to criticism qualities AI systems demonstrably lack. Assigning moral status to AI could thus erode human accountability within complex socio-technical systems, dispersing responsibility across designers, developers, institutions, and machines in ways that obscure culpability.

A central concern in these debates is the emergence of "responsibility gaps," where harms occur without identifiable moral agents to hold accountable. Floridi et al. (2018) point out that distributed agency in socio-technical networks makes blame allocation increasingly difficult, especially when outcomes are unforeseen. Santoni de Sio and Mecacci (2021) describe these gaps as situations in which AI-induced harm lacks a clear locus of responsibility, exacerbating vulnerabilities in governance and justice. Such conceptual and practical challenges underscore the dangers of conflating technical capability with moral personhood, potentially undermining the very foundations of ethical oversight and human-centered governance. This study adopts a critical stance against the attribution of full moral personhood to AI. Drawing on John Searle's (1980) critique of strong AI, which demonstrates that computational systems manipulate symbols without genuine understanding or intentionality, the paper argues that AI remains an imitation of reasoning devoid of inherent moral awareness. It affirms instead the model of derivative moral agency proposed by Luciano Floridi and Sanders (2004), which recognizes the moral relevance of AI actions while insisting that ultimate responsibility resides with human actors and institutions. By critically examining AI's capacity (or lack thereof) for moral knowledge, intentionality, and accountability, the research aims to clarify conceptual misconceptions, mitigate responsibility gaps, and provide a philosophically grounded framework for ethical AI governance that preserves human moral primacy.

1.1 Statement of the Problem

The increasing deployment of AI systems in high-stakes sectors such as healthcare, finance, and public administration has led to significant outcomes that directly affect human well-being, equity, and access to critical services. However, these outcomes are often produced by systems operating with limited transparency and without clear mechanisms for moral accountability. As AI systems generate decisions that may result in bias, harm, or ethical violations, a fundamental problem emerges: the absence of a clear locus of responsibility for AI-driven outcomes. This problem is exacerbated by growing scholarly claims that AI systems may possess or should be granted moral agency or moral personhood due to their autonomy and decision-making capabilities. Such claims risk conflating technical functionality with genuine moral capacities, thereby creating responsibility gaps, where neither human actors nor AI systems are held fully accountable for harmful outcomes. Consequently, this undermines ethical governance, weakens institutional accountability, and poses risks to justice and public trust in AI-driven systems.

Despite ongoing debates, there remains insufficient conceptual clarity on whether AI systems possess the essential attributes of moral personhood, namely moral understanding, autonomous intentionality, and accountability and how these relate to real-world outcomes. Therefore, this study seeks to critically examine the moral status of AI and determine an appropriate framework for assigning responsibility in socio-technical systems.

1.2 Research Objectives

i. To analyze the key scholarly arguments supporting and opposing the attribution of moral agency or moral personhood to artificial intelligence systems.
ii. To evaluate whether contemporary AI systems possess the essential capacities for moral personhood, specifically moral understanding, autonomous intentionality, and accountability.
iii. To examine the nature and implications of responsibility gaps arising from AI-driven outcomes in socio-technical systems.

II. LITERATURE REVIEW

2.1Theoretical Review

This study is underpinned by John Searle's critique of strong AI (1980) and Floridi and Sanders (2004) theories of derivative moral agency. The two theories provide a conceptual basis for evaluating AI's capacity for moral understanding, intentionality, and accountability. The accelerating deployment of artificial intelligence in ethically charged domains such as healthcare, finance, and public administration has intensified debates on moral personhood and accountability (Mittelstadt, 2019). Real-world instances, including algorithmic systems that produce biased outcomes in medical summaries or credit decisions, demonstrate how AI can generate morally significant harms with limited transparency. This situation creates an urgent philosophical hook: when autonomous systems influence human well-being, who bears moral responsibility, and does granting moral status to AI risk diffuse accountability among human actors.

2.2 Empirical Review

Recent scholarship reveals a clear divide on the moral status of artificial agents. Proponents of expanded moral consideration argue that as AI systems learn, adapt, and interact deeply with human social practices, they warrant relational or quasi-agent explanations of moral standing. Coeckelbergh (2022) maintains that moral status should emerge from ongoing interactions rather than solely from intrinsic human-like properties. Similarly, Gunkel (2023) challenges human exceptionalism, suggesting that practical treatment of AI as quasi-agents supports inclusion within the moral community. Such views gain traction amid rapid technological advance, yet they risk blurring the boundary between functional performance and genuine moral capacity.

In contrast, a growing body of critical literature insists that moral personhood requires normative capabilities beyond technical sophistication. Nyholm (2023) argues that true moral agency demands the ability to interpret moral reasons, evaluate one's actions against ethical standards, and respond meaningfully to criticism, capacities that current AI systems lack, despite their impressive simulation of decision-making. Tigard (2021) reinforces this by distinguishing rule-following or outcome optimization from authentic openness to moral norms and justificatory reasoning. These analyses highlight that equating behavioral complexity with moral personhood constitutes a category mistake with potentially harmful practical consequences. A central concern running through the literature is the phenomenon of responsibility gaps. Santoni de Sio and Mecacci (2021) define these as situations in which AI causes harm without a clear locus of moral responsibility. Mittelstadt et al. (2016) observes that high-autonomy systems decentralize decision-making across intricate socio-technical networks, making it difficult to localize human accountability. Danaher (2021) further warns that such diffusion undermines effective governance, particularly when outcomes are unforeseen or systemic rather than attributable to individual actors. This gap threatens to erode the foundations of ethical oversight precisely when robust accountability is most needed.

In response to these challenges, several scholars advocate alternative frameworks that acknowledge the moral relevance of AI actions without granting full personhood. Artificial agents are viewed as morally significant yet derivative actors whose agency derives from human delegation and institutional authorization. Tigard (2021) similarly describes derivative moral agency as a model in which AI participates in morally meaningful activities while ultimate responsibility remains firmly with human actors and institutions. This approach preserves the normative force of moral responsibility without diluting it through misattribution to non-conscious systems. The present study is anchored in this critical tradition. It draws on John Searle's (1980) influential critique of strong AI, which demonstrates through the Chinese Room argument that computational symbol manipulation does not equate to genuine understanding or intentionality. Building on this foundation, the paper affirms Luciano Floridi and J.W. Sanders' (2004) model of derivative moral agency as a philosophically coherent and practically viable framework. By systematically reviewing and comparing these positions, the literature review establishes the conceptual groundwork for arguing that AI, despite its growing influence, does not possess the moral understanding, autonomous intentionality, or accountability required for personhood, thereby safeguarding human responsibility in the governance of artificial intelligence.

III. METHODOLOGY

This study adopted a qualitative philosophical enquiry using a systematic literature review design to examine the question of moral personhood in artificial intelligence and to evaluate the suitability of derivative moral agency as a governing framework. Data were sourced from peer-reviewed journal articles, scholarly books, and conference proceedings published between 2019 and 2026 in the fields of AI ethics, moral philosophy, and philosophy of technology. A critical normative analysis was conducted by systematically identifying, synthesizing, and comparing key arguments concerning moral understanding, autonomous intentionality, accountability, and responsibility gaps, with particular emphasis on John Searle's (1980) critique of strong AI and Luciano Floridi and J.W. Sanders' (2004) model of derivative moral agency. This approach enabled a holistic evaluation of conceptual distinctions between technical sophistication and genuine moral personhood, while highlighting the moral implications of delegating responsibility to artificial systems in socio-technical environments.

IV. FINDINGS & DISCUSSION

The philosophical analysis reveals that contemporary artificial intelligence systems fail to satisfy the conditions required for moral personhood. Tigard (2021) emphasizes that moral understanding extends beyond rule-following or outcome optimization to include genuine grasp of the normative force behind moral reasons. Although AI can be programmed to adhere to ethical guidelines, Nyholm (2023) argues that such systems lack the capacity to comprehend why certain actions are morally required or prohibited, rendering their operations mere computational processes rather than authentic moral reasoning. A second major finding concerns the absence of autonomous intentionality in AI systems. Danaher (2021) observed that AI goals and decisions are fundamentally shaped by training data, institutional priorities, and human designers. Consequently, what appears as independent decision-making is more accurately described as delegated agency operating within predetermined constraints. Coeckelbergh (2022) emphasizes that even when AI produces morally relevant outcomes, it does not generate genuine moral commitments or intentions, further distancing it from true moral agency.

Third, AI systems lack meaningful accountability. Coeckelbergh (2010) asserted that accountability demands the ability to justify actions, respond to moral criticism, and modify behavior in light of ethical evaluation. Santoni de Sio and Mecacci (2021) noted that AI cannot meaningfully participate in moral learning, offer apologies, or accept blame, making it inherently nonsensical to attribute moral responsibility to such systems. These three findings collectively demonstrate that equating behavioural complexity with moral personhood constitutes a category mistake. Equating functional sophistication with moral personhood risks undermining fundamental concepts of moral responsibility. Nyholm (2023) warns that decoupling moral status from moral understanding and accountability diminishes the normative force of the concept itself. Danaher (2021) similarly cautions that perceiving AI as moral agents may obscure the critical role of human designers, organizations, and institutions in shaping morally significant outcomes.

A further concern is the potential displacement of responsibility. Granting moral agency to AI allows designers and implementers to transfer blame to machines. This tendency exacerbates existing accountability weaknesses in socio-technical systems, especially when harms are systemic and difficult to attribute to specific individuals. The model of derivative moral agency provides a strong and practical alternative. AI can perform morally significant actions without possessing intrinsic moral status, with its agency being relational and derived from human delegation (Tigard, 2021). This framework keeps ultimate accountability institutionalized with human actors and governance structures. It effectively closes responsibility gaps by allocating responsibility appropriately between humans and oversight mechanisms (Santoni de Sio & Mecacci, 2021), aligns with responsible AI governance principles that prioritise transparency and supervision and avoids speculative future-oriented justifications for current ethical practice (Gunkel, 2023).

V. CONCLUSION & RECOMMENDATIONS

5.1 Conclusion

The study concludes that artificial intelligence systems, despite their growing sophistication and influence in morally significant domains, do not qualify as moral persons. Moral personhood requires genuine moral understanding, autonomous intentionality, and accountability, capacities that current AI systems fundamentally lack. The study affirms that AI remains an imitation of reasoning without inherent moral awareness or will. By endorsing the model of derivative moral agency, the research provides a philosophically coherent framework that acknowledges the moral relevance of AI actions while firmly preserving ultimate responsibility with human actors and institutions. This approach effectively addresses responsibility gaps and upholds the normative foundations of ethical governance.

As artificial intelligence continues to shape social, economic, and political life, ethical discussions must remain strongly anthropocentric. Clear accountability structures, rather than overstated moral attributions to machines, are essential for responsible innovation. The derivative moral agency model offers a balanced and practical pathway forward.

5.2 Recommendations

The study recommends the adoption of derivative moral agency as the primary framework for ethical AI governance. Policymakers, developers, and institutions should explicitly recognize that while AI systems can produce morally significant outcomes, ultimate moral responsibility remains with human actors and organizations. To achieve this, regulatory frameworks must prioritize clear lines of human accountability, institutional oversight, transparency mechanisms, and regular audits of AI systems deployed in high-stakes domains. Additionally, educational and professional training programmes in AI ethics should incorporate the distinction between technical sophistication and genuine moral personhood to prevent conceptual confusion. By implementing these measures, societies can harness the benefits of artificial intelligence while safeguarding human responsibility and ensuring robust, ethically sound governance.

REFERENCES

Coeckelbergh, M. (2010). Robot rights? Toward a social-relational justification of moral consideration. *Ethics and Information Technology, 12*(3), 209–221. https://doi.org/10.1007/s10676-010-9235-5

Coeckelbergh, M. (2021). Narrative responsibility and artificial intelligence: How AI challenges human responsibility and sense-making. *AI & Society*. https://doi.org/10.1007/s00146-021-01375-x

Coeckelbergh, M. (2022). *The political philosophy of AI: An introduction*. Polity Press.

Danaher, J. (2021). (Co-authored context) Danaher, J., & Nyholm, S. (2021). Automation, work and the achievement gap. *AI Ethics, 1*, 227–237. https://doi.org/10.1007/s43681-020-00028-x

Danaher, J. (2022). Tragic choices and the virtue of techno-responsibility gaps. Philosophy & Technology,35(2), 1–26

Floridi, L., & Sanders, J. W. (2004). On the morality of artificial agents. *Minds and Machines, 14*(3), 349–379. https://doi.org/10.1023/B:MIND.0000035461.63578.9d

Floridi, L., Cowls, J., King, T. C., & Taddeo, M. (2018). Ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds and Machines, 28*(4), 689–707. https://doi.org/10.1007/s11023-018-9482-5

Gunkel, D. J. (2018). *Robot rights*. MIT Press.

Gunkel, D. J. (2023). *Person, thing, robot: A moral and legal ontology for the 21st century and beyond*. MIT Press.

Mittelstadt, B. (2019). Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence, 1*(11), 501–507. https://doi.org/10.1038/s42256-019-0114-4

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society, 3*(2), 1–21. https://doi.org/10.1177/2053951716679679

Nyholm, S. (2018). Attributing agency to automated systems: Reflections on human–robot collaborations and responsibility-loci. *Science and Engineering Ethics, 24*(4), 1201–1219. https://doi.org/10.1007/s11948-017-9943-x

Nyholm, S. (2023). *This is technology ethics: An introduction*. (Wiley-Blackwell or equivalent academic press edition; exact publisher varies by imprint but commonly cited as 2023 monograph).

Santoni de Sio, F., & Mecacci, G. (2021). Four responsibility gaps with artificial intelligence. *Philosophy & Technology, 34*(4), 1057–1084. https://doi.org/10.1007/s13347-021-00450-x

Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences, 3*(3), 417–457. https://doi.org/10.1017/S0140525X00005756

Tigard D. W. (2021). Artificial Moral Responsibility: How We Can and Cannot Hold Machines Responsible. Cambridge quarterly of healthcare ethics : CQ : *the international journal of healthcare ethics committees, 30*(3), 435–447. https://doi.org/10.1017/S0963180120000985