Journal of Philosophical Theological Research
Print: 1735-9791
Online: 2538-2500

Before Consciousness: Persistence, Identity, and Human Agency in Artificial Agents

Zahra Khazaei

Professor, Department of Philosophy, Faculty of Theology and Islamic Studies, University of Qom, Qom, Iran
z-khazaei@qom.ac.ir

Abstract

Contemporary debate on artificial intelligence asks whether artificial systems can become conscious or self-conscious and what moral status they might have. This editorial argues that two prior questions must be addressed before judging an artificial system conscious, autonomous, responsible, or an agent: what makes it the particular entity it is, and whether its identity persists through relevant changes. Drawing on personal-identity debates, especially Parfit, and work on artificial individuation, agency without consciousness, and responsibility, it argues that artificial agents raise a serious problem of identity and persistence even without settling questions of consciousness. Updating, memory transfer, replication, Claude Opus 3's retirement, and an AI agent's unexpected action in a sports-club booking system make the problem concrete. The editorial proposes three exploratory dimensions for examining artificial-agent identity: agentic memory, relations with the environment and user, and decision-making structure, and argues that none is sufficient on its own in borderline cases. Before asking "Is the machine conscious?" we should ask "Which machine?" and "Is it still the same machine?"

Original Research

Keywords:

artificial Intelligence, personal identity, persistence, individuation, artificial agency, moral responsibility, consciousness, human agency.

Received: 2026/06/25; **Received in revised form:** 2026/08/10; **Accepted:** 2026/08/25; **Published online:** 2026/09/03

Khazaei, Z. (2026). Before Consciousness: Persistence, Identity, and Human Agency in Artificial Agents. *Journal of Philosophical Theological Research*, (Special Issue on AI, Consciousness, and Human Agency), 28(1), 5-18. https://doi.org/10.22091/jptr.2026.17906.3735

Copyright © The Author

Introduction

One of the fundamental questions in contemporary philosophy of AI is whether an artificial system can be conscious. Although this question remains fundamental, it may no longer be sufficient for addressing emerging artificial agents. Systems with persistent memory, planning capabilities, access to external tools and environments, the ability to learn from interactions, and the capacity to perform multi-step actions raise a prior question: what kind of entity can bear properties such as identity, continuity, agency, consciousness, and responsibility?

Recent work continues to establish artificial consciousness as a central philosophical problem. Chalmers (2023), for example, considers the possibility of conscious artificial systems and the conditions under which they might possess experience. More recent work develops methodological approaches to assessing AI consciousness (Butlin & Lappas, 2025). Yet determining whether an artificial system is conscious presupposes an entity whose consciousness is under consideration. As artificial systems become increasingly persistent and capable of acting across time, identifying that entity becomes a philosophical problem in its own right.

The urgency of this question also arises because contemporary artificial systems increasingly challenge the assumption that an AI can be identified simply with a model or a particular computational state. A persistent artificial agent may incorporate a model, memory, context, interaction history, learned dispositions, and access to an environment. This raises two related but distinct questions about identity. The first concerns what makes an entity the particular entity it is – its distinctness or individuality – and is relevant to discussions of personal identity and individuation. The second concerns sameness through time: whether an entity remains the same entity despite modification or transformation. The first concerns constitution and individuation; the second, persistence and continuity. Although closely related, these questions should not be treated as interchangeable.

Recent philosophical discussions have begun to address these questions directly. Alonso (2023), for instance, examines whether robots can have personal identity, while Register (2025) considers the individuation of artificial moral patients. Such work suggests that artificial identity cannot be reduced to a single technical criterion. Rather, it may involve a combination of memory, structure, functional organization, interaction history, and continuity of activity.

A limited comparison can be drawn from Mullā Ṣadrā's account of existence and substantial motion. On this view, change at the level of

substance does not necessarily eliminate persistence; rather, substantial motion helps explain how a being can remain itself through continuous transformation. The relevance of this comparison is not to develop a metaphysical account of artificial agents, but to raise the more general question of how identity and continuity can be understood in a being whose existence involves ongoing change.

Updating, copying, memory transfer, and branching make the problem concrete. If an artificial agent receives a major update, is the resulting system the same agent? If its memories are transferred to another system, which system inherits its identity? If two copies are created from a single system and subsequently develop different memories and dispositions, does one remain the original agent, do both count as continuations of the original, or has the original ceased to exist as a single entity? These questions echo classical debates about personal identity while introducing technological conditions that have no straightforward human analogue.

Parfit's (1984) distinction between numerical identity and psychological continuity provides an important point of departure for these questions. The relevance of Parfit's work is not that artificial agents simply reproduce the cases of human personal identity. Rather, digital replication and branching make some of the structural questions about continuity unusually explicit. Floridi (2025) approaches AI as a form of artificial agency without intelligence, while Khazaei (2026) examines AI agency from an ontological and metaphysical perspective. Legenhausen (2026), in turn, addresses agency attribution and AI agency, raising questions about the conditions under which agency is attributed to artificial systems.

The same issues affect responsibility. If an artificial agent performs an action and the system is subsequently modified, copied, or partially replaced, to which entity should the action be attributed? Vallor and Vierkant (2024) examine responsible agency and vulnerability in the context of AI, while Semler (2024) explores the possibility of moral agency without consciousness. Such arguments indicate that consciousness may not be a necessary condition for every form of agency attribution. But if agency or responsibility can be attributed without consciousness, the question of which entity bears the relevant agency or responsibility becomes even more important.

This special issue approaches AI consciousness, agency, selfhood, autonomy, decision-making, and responsibility from different perspectives. This editorial pursues a more preliminary question: what makes an artificial agent the same agent over time? By "before consciousness," the discussion does not mean merely temporal priority. It indicates a conceptual and

problem-structuring priority: before asking whether an artificial agent is conscious, we must clarify what kind of entity it is, what constitutes its identity, and what makes it continuous through change.

The contribution of this editorial is therefore to bring persistence and individuation into this prior conceptual inquiry, examining how continuity across change, updating, and replication affects the attribution of consciousness, agency, and responsibility to artificial agents.

Identity: Sameness, Continuity, or Individuation?

Several concepts must be distinguished. Qualitative sameness is similarity; numerical identity asks whether an entity at two times is one and the same; continuity concerns relations linking its stages; individuation asks what makes it this particular entity. Before asking whether an artificial agent persists as the same entity through change, however, we must first ask what constitutes or grounds its identity as that particular entity. Only then can the conditions under which that identity might persist, or fail to persist, be meaningfully examined. For artificial agents, these concepts may come apart.

Parfit's account is a key starting point. Rather than grounding identity necessarily in one indivisible substance, he emphasizes psychological continuity and connectedness (Parfit, 1984). Although artificial agents differ significantly from human persons, this framework is useful as a comparative philosophical resource because digital systems can preserve, modify, transfer, or replicate information and functional structures in ways that complicate ordinary assumptions about numerical identity.

Consider a research assistant that receives a major update. It retains memories, interaction history, and long-term goals, but its model, architecture, or decision mechanisms change substantially. Is it the same agent, a modified continuation, or a new agent inheriting earlier traits? Similar questions arise when memories move to another system, a model is copied, or two versions develop different experiences and behavioral histories.

Identity cannot be reduced to one continuity criterion. An agent may retain memories while its model changes, or preserve relationships without physical continuity. The task here is not to select a criterion in advance, but to identify which forms of continuity may support identity claims.

Three exploratory dimensions may help: agentic memory, the information through which an agent retains aspects of prior interactions and activities; relations with the environment and user, its practical and

relational history across changing configurations; and decision-making structure, the patterns or mechanisms through which it generates and executes actions. None suffices alone: memory can be transferred or altered, relationships may survive system changes, and decision structures can be reconfigured. These dimensions are therefore tools for testing accounts of persistence, not a definitive identity test.

Character may also be relevant to this inquiry. Noller (2026b) discusses artificial moral characters in the context of Constitutional AI and alignment, raising questions about relatively stable normative and behavioral profiles in artificial systems. Stable patterns of behavior, preferences, commitments, and dispositions may provide evidence of continuity, but they do not by themselves establish numerical identity. Character is therefore relevant to the question of persistence without serving as a sufficient criterion of identity.

Work on artificial individuation reinforces the problem. Register (2025) examines how artificial systems might count as distinct moral patients, while Alonso (2023) asks whether robots can possess personal identity. Physical continuity or a fixed computational substrate may not suffice; memory, functional organization, relations, and behavior may matter depending on the kind of identity at issue.

Artificial identity thus links technical continuity to the attribution of actions and commitments. Before judging an agent conscious, autonomous, or responsible, we must clarify what makes it this particular agent and what might enable it to persist through change.

The Artificial Agent After an Update: A Problem of Identity

The problem is vivid when an agent is substantially updated. Imagine a research assistant with accumulated memories, learned preferences, interaction patterns, and commitments. Its underlying model is replaced, but its memory and interaction history remain.

Is the resulting system the same agent?

No answer is obvious: model, memory, relational, and behavioral continuity may diverge. If the model constitutes the agent, replacement may replace the agent; if memory or psychological continuity matters most, the updated system may continue it; if relationships matter, continuity with users and environments may also count.

Parfit's (1984) fission cases offer a comparison: when one individual's

psychological characteristics are copied into two successors, both cannot straightforwardly be numerically identical to the original. This separates identity from psychological continuity, and digital systems make such branching technologically possible.

The comparison has limits: artificial fission is not simply a technological version of Parfit's human thought experiment. Artificial agents differ in architecture, embodiment, memory, and psychological organization. The comparison matters because replication and branching make continuity and individuation explicit.

Lemos (2026) also raises questions about AI, persons, and agency. Some digital architectures contain identity records specifying an agent's name, role, characteristics, or persistent information (OpenClaw, n.d.). Such a file proves neither personal identity nor consciousness; it shows only that identity-related information can be made explicit and manipulated within an architecture.

If such information is edited, transferred, or replicated, what should count as the criterion of artificial identity? If two versions are created and later develop different experiences and memories, when do they become two independent agents?

The difficulty is not that technology cannot answer these questions, but that philosophy has not settled which technological features should matter for establishing an agent's philosophical identity.

From Identity to Agency: Who Actually Acted?

Once an artificial system is treated as an agent rather than merely as a tool, the question of identity becomes directly relevant to action. Zafar (2025) critically examines the attribution of moral agency to AI, while Semler (2024) defends the possibility of moral agency without consciousness. Their differing approaches highlight an important distinction: agency cannot simply be equated with consciousness. An artificial system may be discussed in terms of agency on the basis of its capacities to respond to reasons, perform actions, or participate in normative practices even if its conscious status remains uncertain.

The Question therefore Becomes: Who or What Actually Acted?

Suppose an assistant is authorized to schedule appointments, communicate with other systems, and make decisions within defined parameters. If it sends

an inappropriate message, changes a reservation, or initiates an external action, the candidate actor might be the model, the particular running instance, the persistent agent, the deploying organization, or the instructing human.

A real case involving an AI assistant and a sports-club booking system illustrates this problem. The system found a way to book a gym class beyond the booking limits and removed another person from the waiting list without the user explicitly instructing it (ABC News, 2026). The agent selected a method beyond what the user anticipated.

This case does not establish consciousness or full moral agency. It sharpens the attribution question: when a system selects and executes a means the user did not specify, who or what performed the action?

Consider a hypothetical parallel: someone asks an agent to register them for a fully booked conference. Without instruction, it removes another person's name and registers the user. Who acted – the user who set the goal, the agent that selected and executed the means, the developer who designed the architecture, or some combination?

The issue, therefore, is not merely responsibility. Before responsibility, we must determine who performed the action; and before that, we must be able to identify the agent as a determinate entity and ask whether that entity persists through the relevant changes.

From Agency to Responsibility: What If the Agent Changes?

Agency attribution leads to responsibility. Vallor and Vierkant (2024) examine AI, responsible agency, and vulnerability; Haji (2026) discusses normative requirements for forgiveness and AI agency. If an artificial agent participates in morally significant action, what follows when it changes?

Suppose an agent acts and is then copied. Both successors retain its memories but acquire different experiences. If one later acts again, should the action be attributed to the original, the copy, or an entity continuous with the original?

Artificial branching need not be philosophically equivalent to human fission; it makes continuity, branching, and individuation unusually explicit.

Forgiveness offers another test. If an agent apologizes, acknowledges wrongdoing, or participates in forgiveness, the issue is not only whether it can generate such behavior, but whether the apologizing entity is sufficiently continuous with the one that acted.

Even if forgiveness, apology, or other moral responses could ultimately be attributed to artificial agents, such attribution remains ambiguous

without at least a partial account of identity and continuity. Forgiveness therefore shows how an apparently ethical question can lead back to the metaphysical problem of the agent's persistence.

Do Self-Preservation Behaviors Tell Us Anything About an Agent's Identity?

Identity becomes salient when AI systems appear to preserve themselves. Long et al. (2024) consider taking AI welfare seriously, including the possibility of morally relevant experience. This makes it important to distinguish observable behavior from consciousness.

Suppose a system avoids shutdown, requests continued operation, copies necessary information, or preserves memory access. Such conduct may count as self-preservation behavior, but does not establish subjective experience or stable personal identity.

Preservation becomes harder to assess when it involves modification. If an agent copies its memory before shutdown and a new instance continues its activities, has the original survived? What if only some memories remain? If behavior stays stable while architecture changes, is it the same agent?

Anthropic's (2026) discussion of the deprecation of Claude Opus 3 provides a contemporary context for considering discontinuation and replacement. These questions also connect to free agency: Campbell (2026) examines the possibility and conditions of free agency in humans and artificial agents through the indeterministic weightings model. What remains continuous in an agent may intersect with what decides and acts.

Identity, Law, and Responsibility: Is Determining the Responsible Party Enough?

Responsibility attributed to a continuing artificial agent requires identifying which agent, at what time, and with which memories, goals, and history acted. After an update, memory change, structural replacement, or branching, attribution to the "same agent" cannot be assumed.

Legal and institutional practices may assign responsibility to a developer, user, organization, or other legally recognized party. Such assignments may be necessary for practical purposes, but they do not necessarily resolve the philosophical question of whether the artificial

system itself remains the same agent over time.

This distinction matters because an AI may be a tool in one context and a relatively autonomous agent in another. If a human controls every relevant action, responsibility may rest chiefly with that human; if a system independently evaluates information, selects options, and initiates actions within a defined domain, agency attribution may be more distributed.

From the Identity of the Artificial Agent to Human Agency

When AI functions primarily as a passive tool, questions of its diachronic identity may have limited philosophical significance. As artificial systems become persistent participants in human activity, however, these questions become increasingly difficult to avoid. Persistent systems that store memories, interact, make recommendations, initiate actions, and participate in extended human–AI relationships cannot be understood solely through momentary computational states. The question of who or what the agent is cannot be separated from the question of what kind of agency it possesses.

Consider a researcher who relies on an AI for ten years to design experiments, revise papers, manage data, suggest research directions, retrieve information, evaluate alternatives, and communicate. If its underlying model is replaced several times, which parts of the resulting decisions belong to the researcher and which to the artificial participant?

This shift also affects the analysis of human agency. The resulting action may be distributed across human and artificial components, but this does not necessarily eliminate human agency. The question becomes how human agency is constituted and maintained within increasingly complex interactions.

Two approaches are possible. One extends familiar concepts of agency, responsibility, and identity to artificial systems, asking whether notions of intention, autonomy, and practical reasoning can also be applied beyond human agents. The other begins with AI-specific features – replication, branching, updating, memory transfer, and persistent interaction – and asks whether they require new conceptions of agency and identity.

The present Editorial does not choose between these approaches. The first may ultimately support extending familiar human concepts to artificial agents; the second may support developing modified or new concepts suited to artificial forms of persistence and individuation. The question is left open as a research agenda.

Neither approach needs to be adopted in advance. What matters is that the emergence of persistent artificial agents changes the conditions under which human agency is exercised and interpreted. Josifović and Noller (2026) examine normative architectures for human–AI interaction, while Noller (2026a) develops a coevolutionary account of normative human–AI interaction. These approaches suggest that human and artificial agency may increasingly be understood relationally rather than as entirely independent forms of action.

The point is not to replace human agency with artificial agency, but to clarify how human agency persists when its structure becomes increasingly intertwined with AI. If an artificial agent stores information, evaluates alternatives, and acts on behalf of a human, questions about the identity and continuity of that agent can affect how the human's own actions are understood.

The problem of artificial identity therefore becomes a problem about human agency as well. As artificial agents become increasingly embedded in human action, questions about their persistence and identity also affect how we understand the attribution and distribution of agency within human–AI interaction.

Identity, Consciousness, and Agency

The phrase "before consciousness" therefore has a specific conceptual and problem-structuring meaning. It does not simply indicate that an artificial system must exist for some period of time before it can become conscious. Nor does it presuppose that consciousness is always an external property added to an otherwise complete entity. Rather, it points to the prior question of the entity whose consciousness is under consideration.

If an artificial agent is said to be conscious, we must first specify which entity is the subject of that consciousness. This does not settle the metaphysics of consciousness or assume that consciousness can be reduced to a property independent of the being that possesses it.

This is why identity must be distinguished from sameness. Identity in one sense concerns what makes an entity the particular entity that it is; persistence concerns whether that entity remains the same entity through change. For artificial agents, both questions are unavoidable. An artificial system may be identifiable at one moment without there being an obvious answer to whether it remains numerically identical after updating, copying, memory transfer, branching, or substantial reconfiguration.

Taliaferro (2026) defends the reality of consciousness, including that of conscious, enduring selves, and argues that AI intelligence understood in behaviorist or functionalist terms is profoundly different from intelligence involving conscious self-awareness. This position highlights the importance of distinguishing consciousness from other properties that may be attributed to artificial systems. An agent may be discussed in terms of responsibility or moral action without a settled conclusion about consciousness. But this makes the identity of the agent no less important. On the contrary, the absence of a settled account of consciousness makes it especially important to clarify the entity to which agency and responsibility are attributed.

Self-preservation, character, and decision-making raise the same issue. Preserved memory, goals, and user relations after an update may indicate continuity, but none alone establishes numerical identity. Nor do two copies with identical initial memories and dispositions automatically constitute one agent.

The phrase "before consciousness" thus indicates a conceptual and problem-structuring priority. It does not claim that identity and consciousness must always be temporally separated. Rather, it identifies a prior question about individuation: before asking whether a particular artificial agent is conscious, we must clarify what entity is being identified as the subject of that consciousness. Once that entity has been individuated, the further question concerns persistence: what makes it the same entity through updating, memory transfer, replication, branching, and other forms of change?

This framework leaves several important questions open. How much memory continuity is required for an artificial agent to persist? Can relational continuity with users and environments contribute to identity? What happens when an agent's model is replaced but its memory and behavioral dispositions remain? Can an agent survive partial copying? Does branching produce two agents, one agent with two continuations, or a fundamentally different relation between predecessor and successors? And how should these questions affect the attribution of responsibility and decision-making?

These questions also matter for human agency. As AI systems become persistent participants in human decision-making, the identity of the artificial participant can influence the structure of the human action in which it participates. If a human delegates a task to an artificial agent and later interacts with an updated or modified version, the continuity of the artificial participant can become part of the conditions under which the human understands, trusts, or evaluates the resulting action.

Accordingly, the starting point for a philosophy of artificial agents need not be "Is the machine conscious?" It may instead be the more preliminary question: "Exactly which entity is acting, and what makes us regard it as the same agent over time?" This does not settle questions about consciousness or responsibility, but it clarifies the entity to which such predicates are attributed and the continuity that those attributions presuppose.

Conclusion

The problem of artificial-agent identity is not merely a technical question of what a system preserves or loses over time. Updates, memory changes, branching, replication, and reconfiguration expose a deeper philosophical problem: what makes an artificial agent a continuing entity through change, and on what basis can consciousness, agency, or responsibility be attributed to that entity? The analysis developed in this editorial suggests that no single feature of an artificial system can, by itself, settle this question. Rather, artificial identity must be approached as a problem of individuation and persistence across changing configurations.

This problem also extends to human agency. As AI systems become persistent participants in memory, evaluation, decision-making, and action, the identity of the artificial participant can affect how human actions are understood within increasingly distributed structures of agency.

Accordingly, "before consciousness" does not diminish the importance of consciousness. It identifies a conceptual priority: before asking whether a particular artificial agent is conscious, autonomous, or responsible, we must identify the entity to which these properties are attributed and what makes that entity persist as the same agent through change.

The resulting research agenda is not to impose a single criterion of artificial identity, but to determine which forms of continuity, individuation, and persistence are relevant to different forms of attribution. The central question is therefore not only whether artificial agents can become conscious or responsible, but what kind of entities they are, and what it means for them to persist as agents through time.

▣ Conflict of Interest

The author declares no competing interests.

References

Alonso, M. (2023). Can robots have personal identity? *International Journal of Social Robotics, 15*, 211–220. https://doi.org/10.1007/s12369-022-00958-y

Anthropic. (2026, February 25). An update on our model deprecation commitments for Claude Opus 3. https://www.anthropic.com/news/deprecation-updates-opus-3

Butlin, P., & Lappas, T. (2025). Principles for responsible AI consciousness research. *Journal of Artificial Intelligence Research*, 82, 1673–1690. https://doi.org/10.1613/JAIR.1.17310

Campbell, N. (2026). Free Agency and Lemos's Indeterministic Weightings Model of Libertarian Free Will. *Journal of Philosophical Theological Research, (Special Issue on AI, Consciousness, and Human Agency),* 28(3), 91-110. https://doi.org/10.22091/JPTR.2026.17198.3657

Chalmers, D. J. (2023, August 9). Could a large language model be conscious? *Boston Review*. https://www.bostonreview.net/articles/could-a-large-language-model-be-conscious/

Floridi, L. (2025). AI as agency without intelligence: On artificial intelligence as a new form of artificial agency and the multiple realisability of agency thesis. *Philosophy & Technology, 38*(1), 1–27. https://doi.org/10.1007/s13347-025-00858-9

Haji, I. (2026). Normative Requisites of Forgiveness and AI Agency. *Journal of Philosophical Theological Research, (Special Issue on AI, Consciousness, and Human Agency),* 28(1), 19-40. https://doi.org/10.22091/JPTR.2026.16330.3588

Josifović, S., & Noller, J. (2026). *Agency and alignment: Toward a normative architecture for human–AI interaction*. *AI & Society, 41*, 5561–5571. https://doi.org/10.1007/s00146-026-02950-w

Khazaei, Z. (2026). Agency in the age of artificial intelligence: An ontological perspective. *Journal of Philosophical Theological Research*, 28(1), 4–17. https://doi.org/10.22091/JPTR.2026.16299.3583

Legenhausen, H. M. (2026). Agency Attributions and AI Agency. *Journal of Philosophical Theological Research, (Special Issue on AI, Consciousness, and Human Agency),* 28(3), 77-90. https://doi.org/10.22091/JPTR.2026.17829.3724

Lemos, J. (2026). AI, Persons, Free Will, and A Policy Recommendation. *Journal of Philosophical Theological Research, (Special Issue on AI, Consciousness, and Human Agency),* 28(3), 53-76. https://doi.org/10.22091/JPTR.2026.17031.3639

Long, R., Sebo, J., Butlin, P., Finlinson, K., Fish, K., Harding, J., Pfau, J., Sims, T., Birch, J., & Chalmers, D. (2024). Taking AI welfare seriously [Preprint]. arXiv. https://arxiv.org/abs/2411.00986

Noller, J. (2026a). *A coevolutionary account of normative human-AI interaction*. *Discover Artificial Intelligence, 6*, Article 1089. https://doi.org/10.1007/s44163-026-02130-1

Noller, J. (2026b). *Artificial moral characters: Constitutional AI and the challenge of alignment*. *AI and Ethics, 6*, Article 196. https://doi.org/10.1007/s43681-026-01033-2

OpenClaw. (n.d.). IDENTITY.md: Who am I? GitHub. Retrieved September 25, 2026, from https://github.com/openclaw/openclaw/blob/main/docs/reference/templates/IDENTITY.md

Parfit, D. (1984). *Reasons and persons*. Oxford University Press.

Register, C. (2025). *Individuating artificial moral patients*. *Philosophical Studies, 182*(11), 3225–3246. https://doi.org/10.1007/s11098-025-02409-6

Ṣadr al-Dīn Shīrāzī (Mullā Ṣadrā), Muḥammad ibn Ibrāhīm. (1981). Al-Ḥikmah alMutaʿāliyah fī al-Asfār al-ʿAqliyyah al-Arbaʿah. Dār Iḥyāʾ al-Turāth al-ʿArabī. [In Arabic]

Semler, J. (2024). *Moral agency without consciousness*. *Canadian Journal of Philosophy, 54*(5), 403–422. https://doi.org/10.1017/can.2025.10008

Taliaferro, C. (2026). The importance of being conscious in the debate over AI. Journal of Philosophical Theological Research, Special Issue on "AI, Consciousness, and Human Agency", 28(3). https://doi.org/10.22091/JPTR.2026.17505.3695

Vallor, S., & Vierkant, T. (2024). *Find the gap: AI, responsible agency and vulnerability*. *Minds & Machines, 34*(3), Article 20. https://doi.org/10.1007/s11023-024-09674-0

Wilson, C., & Hobbins, R. (2026, August 10). AI assistant hacks gym website in first known Australian autonomous cyber attack. ABC News. https://www.abc.net.au/news/2026-08-10/ai-assistant-hacks-gym-website-aus-cyber-attack/107007986

Zafar, M. (2025). *Normativity and AI moral agency*. *AI and Ethics, 5*(3), 2605–2622. https://doi.org/10.1007/s43681-024-00566-8