Journal of Philosophical Theological Research

The Quarterly Academic Journal of University of Qom

JPTR

Vol. 28 • No. 3 • 2026 • Issue 109

Special Issue on

AI, Consciousness, and Human Agency

Publisher: University of Qom

Managing Director: Mohammad Zabihi

Editor-in-Chief: Zahra Khazaei

Editorial Board:

Edward Wierenga (Emiritus Professor of Philosophy, Department of Philosophy, Rochester University, United States), **Vincent Brümmer** (Professor, University of Utrecht and Dean of the Theological Faculty, Netherland), **Stephen R. Palmquist** (Professor, Hong Kong Baptist University), **Mahmoud Morvarid** (Associate Professor, Institute for Research in Fundamental Sciences (IPM)., Iran), **Reinhard Hesse** (Professor, University of Education Freiburg), **Mohsen Javadi** (Professor, University of Qom, Iran), **Seyed Mostafa Mohaghegh Damad** (Professor, Shahid Beheshti University, Iran), **Nancey Murphy** (Professor, PhD. Philosophy, ThD. Christian theology, Senior Professor of Christian Philosophy, Fuller Theological Seminary, USA), **Mohammad Zabihi** (Professor, Univerisity of Qom, Iran), **Einollah Khademi** (Professor, Shahid Rajaee Teacher Training University, Iran), **Zahra Khazaei** (Professor, University of Qom, Iran), **Hamidreza Ayatollahy** (Professor, Allameh Tabatabaii University, Iran), **Jafar Shanazari** (Associate professor, University of Isfahan, Iran), **Robert Kane** (Distinguished Professor, University of Texas, USA), **Ishtiyaque Haji** (Professor, University of Calgary, Canada), **Charles Taliaferro** (emeritus Distinguished Emiritus Professor of Philosophy, St. Olaf College, USA), **Roger Crisp** (Professor of Philosophy, University of Oxford, UK), **Henk bakker** (Professor of Religion and Theology, Vrije Universiteit Amsterdam, Netherlands), **John Greco** (Professor, Georgetown University,, Washington D. C., US)

Executive Director: Mohammad Baqer Mirsane

Editor: Mohammad Heydarpour

Layout: S. M. Hoseini

Journal of Philosophical Theological Research (JPTR) has been indexed in following databases:

Scopus | Philosopher's Index | EBSCO | ProQuest | Ovid | PhilPapers | Atla | ISC | Index Copernicus | DOAJ | Ulrich | J-Gate | Advanced Sciences Index (ASI) | ROAD | Scientific Journal Impact Factor (SJIF) | Scimago | DRJI | International Innovative Journal Impact Factor (IIJIF) | Universal Impact Factor | I2OR | General Impact Factor | Cosmos Impact Factor | Scientific World Index Journal (SWIJ) | Academic Resource Index | Google Scholar | WorldCat | Citefactor | ijifactor | Europub | esjindex (Eurasian scientific journal index) | Scientific indexing Services (SIS) | Academic Keys | Ricest | Magiran | SID | Noormags | Civilica | Base.

Journal of Philosophical Theological Research

is a product of the joint activity of the University of Qom and

The Iranian Association for Philosophy of Religion

Address: University of Qom, Alghadir Blvd., Qom, Iran

Tel: +98-25-32103360-1 Fax: +98-25-32103360 Po. Box: 37161- 46611

Print ISSN: 9791-1735 | Online ISSN: 2538-2500

http://pfk.qom.ac.ir | **jptr@qom. ac. ir**

Instructions to Authors

1. The manuscript should not have been published before and it is not under consideration for publication anywhere else.
2. The length of the articles should not exceed 8000 words.
3. The title page should include: the name (s) of the authors, the academic degree, the field of study, the affiliation (s) of the author (s), i. e. institution, (department), city, (state), country, an active email address (If address information is provided with the affiliation (s) it will also be published)and the phone number of the author (s). Also, the corresponding author -if the authors are more than one person– should be specified.

   * Acknowledgments of people, grants, funds, etc. should be placed in a separate section on the title page. The names of funding organizations should be written in full.

4. The length of the abstract should be between 200 to 250 words.
5. Please provide 5 to 7 keywords which can be used for indexing purposes.
6. Abstract should include the objectives, methods and results of the research.
7. The article must contain a title, an extended abstract in English, an introduction, the main text, the conclusion and the list of references.

   • The main text file of the article should not contain the name of the author (s).

   • Please do not provide references directly or indirectly to the author (s) in the text and footnotes.

   • Proper names, English terms and explanatory references should be cited in the footnotes.

8. The manuscript should be submitted in MS-Word, with "BZar" font for Persian articles and with "Times New Roman" for English ones.

   * The size of the font for the main text should be 12 and the footnotes should be 9.

   * The English abstract should be typed with "Times New Roman" font sized 11 for text and 10 for bibliography.

9. The articles that are accepted can cite thanks to individuals or organizations that have funded the research before the list of references.
10. Author (s) should follow the latest edition of APA style in referencing.

    • References within the text should be written in parenthesis and in the following fashion: (surname of the author, the year of publication and the page number).

    • In case you have provide a reference to an author that has two books which have been published in the same year, you should distinguish them by using the words "a" and "b" after the year of publication.

    • References at the end of text should be sorted alphabetically and based on the following order:

    • **Article:** author's surname, author's name. (the year of publication). The title of article. *the title of Journal* (italic). the volume. Number. the page numbers.

    McDowell, John J. (1979). *Virtue and Reason*. The Monist 62 (3), 331- 350

    • **Book:** author's surname, author's name. (year of publication). *title of book* (italic). the translator/ editor/ collector. the place of publication. and the publisher.

    Shirazi, S. M., (Mullā Ṣadrā). (1981). *al-Shawahid al-rububiyyah fi manahij al-sulukiyyah*. Sayyid Jalal al-Din Ashtiyani (ed.) together with Sabzevārī's commentary. Mashhad: Markaze Nashre Daneshgahi.

    • **Chapter of a book:** author's surname, author's name. (year of publication). the title of the article. the name of the collector or editor. the title of the reference. the page number. the place of publication: the publisher.

    Zagzebski, L. (2001). Must Knowers be Agent? in A. Fairweather & L. Zagzebski (Eds.). Vitue Epistemology: Essays on Epistemic Virtue and Responsibility (pp. 142-157). Oxford: Oxford University Press.

    • **Online resource:** author's surname, author's name. the title of the article. the address of the website (italic). and the date that the article has been received.

    Taliaferro, Ch. (2013). Philosophy of Religion. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Summer 2013). Retrieved from https://plato. stanford. edu/archives/Sum2013/entries/ Philosophy of Religion

11. **Responsibility of the Article**

    • Authors (s) are solely responsible for the accuracy of the article. Therefore, the articles do not necessarily reflect the views of the editorial board.

    • The editorial board of journal has the right to accept, reject and edit the article.

    • The journal rejects articles that contain elements that are plagiarized. Therefore, if an author uses someone else's research achievement or a part of his/her own published researches without citing the source, the article will be rejected. • JPTR handles the plagiarism and duplication with "Samimnoor" system for Persian articles and "iThenticate" for English articles.

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

Content

■ Editorial

Before Consciousness: Persistence, Identity, and Human Agency in Artificial Agents ..5
Zahra Khazaei

■ Research Articles

Normative Requisites of Forgiveness and AI Agency 19
Ishtiyaque Haji

The Importance of Being Conscious in the Debate over AI..... 41
Charles Taliaferro

Original Research

AI, Persons, Free Will, and a Policy Recommendation 53
John Lemos

Agency Attributions and AI Agency 77
Hajj Muhammad Legenhausen

Free Agency and Lemos's Indeterministic Weightings Model of Libertarian Free Will 91
Neil Campbell

Artificial Intelligence, Consciousness, Transcendence: A Thoughtful Overview between Present and Future 109
Paolo Di Sia

Ethical Authorship in the Age of Generative AI: A Digital Humanist Framework for Scholarly Accountability 125
Reza Rezvani | A. Mehdi Riazi

Consciousness: Human, Artificial, or Both? 155
Viktor J. Tóth

Narrating and Judging: Meta-Qur'anic Speech in Sūrat Ash-Shu'arā' and the Limits of AI 177
Abdelmadjid Ben Habib

A Critical Evaluation of the Fitting-Recipient Argument for the AI Ensoulment Hypothesis 205
Reza Dargahifar

From the Rational Requirement and Ontological Reconstitution of Human Agency to More Onerous Duties 233
Muhammad Heydarpour | Ali Razizadeh

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

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

Normative Requisites of Forgiveness and AI Agency

Ishtiyaque Haji

Professor of Philosophy, Department of Philosophy, University of Calgary, Canada,
ihaji@ucalgary.ca

Abstract

Regarding forgiveness, there is disagreement on the following "contrast pairs." (i) The victim (the person who forgives) forgives the offender (the person who is forgiven) for doing something only if it is morally wrong for the offender to do this thing; or even if she has done no wrong, the victim forgives the offender only if she takes herself to have done wrong. (ii) The victim forgives the offender for doing something only if she is blameworthy for doing this thing; or even if not blameworthy, the victim forgives the offender only if she takes herself to be blameworthy for doing this thing. In this paper, first, I appeal to "Frankfurt examples" to argue that the belief options – the second members – of each contrast pair are more tenable. Second, I discuss why even these belief options threaten forgiveness if determinism is true in virtue of determinism's imperiling wrongness. Third, I broach the issue of whether artificial intelligence agents ("AI agents") can be apt candidates of forgiveness. I synopsize conditions AI agents must satisfy to be fitting subjects of forgiveness, emphasizing agency requisites for blameworthiness and wrongdoing. The summary's principal aim is to animate further inquiry into whether AI agents can be agents of the sort that forgiveness presupposes.

Original Research

Keywords:

artificial intelligence agency, blameworthiness, Frankfurt examples, forgiveness, wrongness.

Received: 2026/02/12; **Received in revised form:** 2026/04/16; **Accepted:** 2026/05/12; **Published online:** 2026/06/05

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

Copyright © The Author

Introduction

Assuming that there are or can be artificial intelligence agents ("AI agents") who are minimal agents – entities capable of performing intentional mental or bodily actions – can AI agents be apt candidates of forgiveness? Answering this question requires multifaceted inquiry into, among other things, whether AI agents can do moral wrong, have beliefs about wrongdoing, be blameworthy for some of their actions, or have beliefs about whether they are blameworthy for their actions. This is because, reasonably, there are conceptual connections among forgiveness, wrongness or belief in wrongness, and blameworthiness or belief in blameworthiness. In this paper I first explore these putative connections. I then briefly summarize a non-exhaustive list of the preconditions AI agents must satisfy to be appropriate candidates of forgiveness, emphasizing agency requisites for blameworthiness and wrongdoing. The primary aim of this synopsis is to motivate further inquiry into whether, in addition to being agents in the minimal sense, AI agents can be the type of augmented agent that forgiveness presupposes.

Forgiveness and Dual Contrast Pairs

Consider these "contrast pairs" regarding putative requirements of forgiveness:

Wrongness Contrast Pair

(i) Objective Member: The victim (the person who forgives) forgives the offender (the person who is forgiven) for doing something only if it is morally wrong for the offender to do this thing. This requirement may be coupled with the additional one that the victim believes that the offender has done wrong.[1]

(ii) Subjective Member: The victim forgives the offender for doing something only if the offender believes that it is morally wrong for her to do this thing and the victim believes that the offender takes herself to have done wrong.

The subjective members may be qualified in various ways, for example, regarding (i), one may venture that the offender justifiably believes that she has done wrong. Such qualifications play no substantial role in what is to come

1. As Margaret Holmgren explains, "Where there is no wrongdoing, there is nothing for the injured person to forgive, although there may be much for her to accept….Forgiveness is an issue only when a moral agent commits an offense, without justification, and in the absence of an exculpating excuse" (2012, 35). (See, e.g., also Murphy & Hampton, 1988, 20; Holmgren, 1993; Hughes, 1993; Hieronymi, 2001, 546; Nelkin, 2013, 175; Allais, 2008; 2014, 41).

because any such qualified view entails the unadorned belief view.

Blameworthiness Contrast Pair

(i) Objective Member: The victim forgives the offender for doing something only if the offender is morally blameworthy for this thing. If true, it may be proposed that this precondition goes together with the provision that the victim believes that the offender is blameworthy.[1]

(ii) Subjective Member: The victim forgives the offender for doing something only if the offender believes that she is blameworthy for this thing, and the victim believes that the offender takes herself to be blameworthy for this thing.

Again, (ii) may be restricted in certain ways. For instance, it may be advocated that the victim nonculpably believe that the offender is blameworthy. Such qualifications will be largely ignored because they are inconsequential to the issues to be discussed below.

These pairs may compete with others. With the first, one may submit that the offender can be forgiven for something she did that is bad but not impermissible.[2] As for the objective member of the second pair, determinism is the thesis that at any instant exactly one possible future is compatible with the state of the universe at that instant and the laws of nature (van Inwagen, 1983, p. 3). One may be tempted to defend the view that even if determinism undermines blameworthiness, determinism does not preclude forgiveness.[3] In this paper, I disregard these other options. My primary focus is to ascertain which member of each "contrast pair" has the upper hand. The winner of each pair will have to contend with rivals. In what follows, first, I appeal to Frankfurt examples to support the subjective variants of each pair. Then I explain why there are reasons to believe that determinism threatens these variants largely owing to determinism's, at least prima facie, imperiling moral

1. Dana Nelkin contends that when "one forgives an offense, one is committed to the proposition that the offender acted freely and was responsible for the offense" (Nelkin, 2013, p. 183). Similarly, Brandon Warmke and Michael McKenna emphasize that "directed forgiveness presumes that the person forgiven is indeed blameworthy, and that the victim had blamed the person previously, or at least held her morally responsible and blameworthy" (Warmke & McKenna, 2013, 205). Warmke also remarks that forgiveness "does not change the fact that one did something for which one is guilty and is liable to punishment by the appropriate moral, civil or even divine authorities" (2016, 13).

2. I use "wrong" and "impermissible" interchangeably.

3. See, e.g., Pereboom, 2001; 2014. Pereboom has recently defended the view that forgiveness presupposes blameworthiness ("protest blameworthiness") of a consequentialist variety (2021, pp. 103-104).

wrongness. Finally, I revert to the issue of whether AI agents can be forgiven.

Forgiveness and Impermissibility

Frankfurt examples provide support for the wrongness contrast pair's subjective member that forgiveness requires belief in wrongdoing and not wrongdoing *per se*. Typically, these examples have been invoked to overturn the principle of alternate possibilities (*PAP-R*): agents are morally responsible for what they have done only if they could have done otherwise. These examples may be divided into two stages. In stage 1 of case *Forgiveness,* Laleh ridicules Arman for questionable reasons. In this stage, she could have refrained from deriding Arman, and it is impermissible for her to deride him, and she is blameworthy for doing so. In stage 2, a "rerun" of stage 1, a counterfactual intervener – Omid – who fails to show his hand, would have forced Laleh to ridicule Arman had she displayed an involuntary sign that she was about to refrain. But Laleh acts no differently in this stage than in the former. It looks like she is blameworthy for her relevant behavior in stage 2 even though she could not have done otherwise. Hence, Frankfurt partisans conclude, *PAP-R* is false (see Frankfurt, 1969).

Add to *Forgiveness* that Arman gets wind of Laleh's untoward behavior (in stage 2), and he forgives her. Without good reason to believe otherwise, it appears that there is nothing suspect about this instance of forgiveness.[1] If you agree, you have reason to discard the objective view that to be forgiven for something, it is impermissible for the offender to do it. Why so? Although Arman, as assumed, forgives Laleh for ridiculing him, in stage 2 it is not impermissible for Laleh to ridicule him because no one can do wrong unless one could have done otherwise, and in stage 2 Laleh could not have refrained from deriding Arman.

The view that impermissibility requires freedom to do otherwise rests on two credible principles:

> *Ought Implies Can* (*OIC*): If it is obligatory for you to do *A*, then you can do *A* (and if it is obligatory for you not to do *A*, then you can refrain from doing *A*).[2]

1. One may propose that, here, one does forgive but the forgiveness is not appropriate or fitting. But then the relevant issue concerning wrongness and forgiveness should be taken to address appropriate forgiveness.

2. E.g., Feldman, 1986; Zimmerman, 1996; Haji, 2016; Vranas, 2007; Littlejohn, 2012. I omit temporal indices here. I take the agent- and time-relativized canonical form of obligation

Just as you cannot be responsible for something unless you are free with respect to it and, thus, you can do it, nothing can be obligatory for you unless you are free regarding it and, so, you are able to do it.

> *Equivalence*: It is obligatory for you to refrain from doing *A* if, and only if, it is impermissible for you to do *A*.

In the estimation of many, this "ought not" is equivalent to "impermissible" principle has the status of a morally deontic axiom or theorem.[1]

> *OIC* and *Equivalence* jointly entail:
>
> *Impermissibility-Implies-Can-Avoid* (*IAvoid*): If it is impermissible for you to do *A*, then you can avoid doing *A*.

It is not impermissible for Laleh to ridicule Arman since she cannot avoid ridiculing him in stage 2. However, Arman forgives her for her behavior. Hence, it is not a precondition of forgiveness that the offender has done wrong. If this precondition is false, it appears that there is little reason to accept the corollary that forgiveness presupposes that the victim believes that the offender has done wrong. In short, the wrongness contrast pair's objective member is false.

A caveat is in order. Suppose you have initial (or "pretheoretical") inclinations regarding which member of each contrast pair under scrutiny is true. For example, it might be that you initially judge that forgiveness surely requires that the offender does wrong. Then you may be drawn to claiming that in Frankfurt examples the victim does not forgive because the offender's action – in *Forgiveness*, Laleh's ridiculing Arman – is not impermissible for her. A relevant comparison is that you may antecedently judge that *PAP-Blame* – the principle that blameworthiness requires freedom to do otherwise – is true. Then Frankfurt examples won't move you to renounce *PAP-Blame*; if you are already committed to *PAP-Blame*, you should insist that in *Forgiveness* Laleh is not blameworthy for ridiculing Arman in stage 2. Similarly, if you antecedently judge that forgiveness presupposes blameworthiness, and blameworthiness requires freedom to do otherwise, Frankfurt examples won't convince you that the victim forgives the offender in Frankfurt cases. However, this would not be the proper way to proceed if the very inquiry concerns which member of each contrast pair is true. On pain

---

statements to be this: at *t*, it is obligatory for *S* to do *A* at *t** (where *t** may be *t* or later than *t*).

1. E.g., Belzer, 1998; Hilpinen & McNamara, 2013, 43; McNamara, 2019; Rönnedal, 2009, 28-29. On *Equivalence*, see, also, Pereboom, 2001, 146-147; and Haji, 2002, 52-53.

of begging the question in one direction or other, it should not be these antecedent judgments that decisively inform your pertinent verdicts regarding forgiveness and impermissibility or forgiveness and blameworthiness in Frankfurt examples.

I now address two objections to this appeal to Frankfurt examples against the wrongness contrast pair's objective member. The first is that its not being impermissible for Laleh to deride Arman in stage 2 yields the absurd result that it is permissible for her to engage in this behavior. However, the inference from "not impermissible" to "permissible" is unwarranted. To see this, imagine that in an amended Frankfurt-like example, Fatemeh is not blameworthy for unjustifiably rebuffing Azad in stage 2 because Omid forces Fatemeh's hand; Fatemeh fails to exercise *responsibility-level control* when she snubs Azad. It would be a mistake to conclude that since Fatemeh is not blameworthy, she is either praiseworthy or "neutrally" responsible for her rebuffing in stage 2. Without the control that responsibility demands, she would not be responsible at all. Analogously, should Laleh fail to exercise *obligation-level control* in ridiculing Arman, it is neither obligatory nor permissible for her to ridicule Arman, in addition to its not being impermissible for her to ridicule him. Rather, Laleh's act of ridiculing Arman is amoral for her: ridiculing Arman is not permissible, impermissible, or obligatory for her.

My position here rests on the principle, *Obligation/Freedom*, that obligation requires control. This principle is homologous to the well accepted principle, *Responsibility/Freedom*, that responsibility requires control. You may intentionally perform an action but in so doing, lack obligation- or prohibition-relevant control in performing it. This action which would otherwise be obligatory (or impermissible) for you is neither. So, what may be said to support *Obligation/Freedom* and, hence, to reject the view that actions are either permissible or impermissible for any moral agent? What may be offered to bolster the view that there is such a category as the *amoral* for moral agents? Although I accept *Responsibility/Freedom* I have no direct argument for it (that is, I have no argument with this principle as its conclusion), nor have I seen any plausible *direct* argument for it. I endorse *Responsibility/Freedom* because it has historical backing, it is intuitively plausible, and reflection on a range of cases persuades me that responsibility requires freedom. I accede to *Obligation/Freedom* for similar reasons: it has historical precedent, it is intuitively credible, attention to relevant cases convinces me that it is reasonable, and accepting it, along with other apt propositions, enables me to achieve reflective equilibrium regarding an overall account of obligation (see Haji, 2019). Here is one sort of case I find especially revealing. Grant that Laleh could not have done otherwise in *Forgiveness*.

Then it isn't impermissible for her to ridicule Arman in stage 2 (given *OIC* and *Equivalence*). I've proposed that it's not impermissible for her to ridicule Arman because she could not have done otherwise.

Suppose one claims that since it is not impermissible for her to ridicule Arman, it must be permissible for her because an act is either permissible or impermissible for any moral agent. First, this result is highly counterintuitive. (To drive home this contention, imagine that the action Laleh cannot avoid performing in stage 2 is pulling the trigger (thereby ensuring that an innocent person is killed)). Second, if you think this action is permissible, you are committed to this asymmetry: Whereas impermissibility requires freedom (freedom to do otherwise), permissibility does not. But I can't think of a plausible reason for the asymmetry in the control requirements of impermissibility and permissibility. Nor should we expect this sort of asymmetry, given the intimate conceptual connection between obligation and impermissibility that *Equivalence* exposes. Elaborating, *A* is optional for you if, and only if, both *A* and not-*A* (refraining from doing *A*) are permissible for you. If it is permissible for you to do *A*, then either it is obligatory for you to do *A*, or it is optional for you to do *A*. Assume some act, *B*, is impermissible for you. Then, you can do otherwise than *B* (since impermissibility requires avoidability), and you have a negative obligation concerning *B* – you ought not to do *B* – since "ought not" is equivalent to "impermissible." Barring strong reason to think otherwise, the freedom requirements of an obligation to do something (a positive obligation) and an obligation to refrain from doing something (a negative obligation or a prohibition) are symmetric. Now suppose, first, *A* is permissible for you because *A* is obligatory for you. Then, again, in the absence of reasons to believe otherwise, why think that in this sort of case, the freedom requirements of permissibility differ from those of obligation, and so differ from those of prohibition or impermissibility on the supposition that negative and positive obligations have symmetric freedom requirements? Second, if in this sort of case, the freedom requirements of permissibility don't diverge from those of impermissibility, why believe that we should expect a divergence in these requirements in the remaining cases in which *A* is permissible for you because *A* is optional for you? The foregoing are not decisive reasons for accepting *Obligation/Freedom* and the proposal that the freedom requirements of permissibility, impermissibility, and obligation are symmetric, but they are not without merit either.

I next turn to a second objection to the view that in *Forgiveness* it's not impermissible for Laleh to ridicule Arman in stage 2. The general worry is that there are different kinds of ability, and if different kinds of ability track different kinds of control, couldn't Frankfurt cases be a reason to *reject* a

particular interpretation of *OIC* that appeals to a sort of ability (specific ability to do otherwise) that Laleh ostensibly lacks in stage 2, but not to reject the kind of control that blameworthiness requires? In response, distinguish between two broad classes of ability, specific and general ability. To illustrate the difference, Shirin, on the road to the market, cannot steal any dates because along the way there are no dates to steal. Although she lacks the specific ability to steal, she has the general ability; even though she has no opportunity to steal, she retains her thieving talents and skills. Entering the market, Shirin plies her trade: she pilfers some dates, and in so doing exercises her specific ability to steal.[1] Informally, and appealing to a different example, you have the general ability to ride a bike insofar as you have the intrinsic powers – balance, motor control, skills and so forth – to ride. Your ability is increasingly specific to the extent that you also have the agreeable state – the physiological and psychological capacities – to ride (see Clarke, 2015; Vihvelin, 2004). Your ability is more specific still to the extent that you additionally have the opportunity – access to a bike and window for use – to ride. Roughly, you have the specific ability to ride a bike if, and only if, you have the intrinsic powers, are in the agreeable state, and have opportunity to ride.

Revisiting *Forgiveness*, with Omid in the wings, and excluding other problems with Frankfurt examples, independently of confirmation by an analysis of specific ability, it's plausible that although Laleh has the general ability to refrain from ridiculing Arman in Stage 2, she lacks the specific ability to refrain.[2] Regarding the objection under scrutiny – that the control requirements of blameworthiness and obligation (or impermissibility) track different sorts of ability – maybe the objector will propose, first, that Frankfurt cases like *Forgiveness* reveal that in stage 2 Laleh lacks the specific ability to refrain from ridiculing Arman but not the specific ability to ridicule Arman. In addition, the objector may insist that blameworthiness for an action requires that one have the specific ability to perform the action but not the specific ability to refrain as Frankfurt examples show. In contrast, obligation requires general and not specific ability or at least some kind of ability that is much more akin to general ability and much less like specific ability. Laleh retains the general ability both to ridicule and to refrain from ridiculing Arman even in stage 2 with the counterfactual intervener on duty. Hence, there is no bar to maintaining that in stage 2 it *is* impermissible for Laleh to ridicule Arman.

---

1. For more on the distinction between specific and general ability, see, e.g., Haji, 2016, p. 213; Hebert 2016.

2. One may propose that even in stage 2, Jones has the *specific* ability to refrain from stealing. See, e.g., Nelkin 2011, 66-71; and Vihvelin 2013, 195-196. For responses to Nelkin and Vihvelin, see Haji 2019, 190-200.

In reply, if there is a requirement of alternate possibilities for obligation and impermissibility, it is implausible that the requisite ability to do otherwise is general and not specific. Proponents of *OIC* cannot sensibly be taken to propose that an obligation to do something presupposes only the general and not the specific ability to do it. Imagine you are on lifeguard duty and a child in the ocean is in serious trouble. You cannot save the child – you don't have the specific ability to save the child – because sharks in the water between you and the child would eat you if you attempted to save the child. Although you have the general ability to save the child – you have the skills and know-how to save – you lack the opportunity to save. Hence, it would be unreasonable to insist that you morally ought to save the child even though you cannot. If you ought to do something, you can do it, and "can" here expresses specific ability. Excluding special reason to believe otherwise, if "ought" and "impermissibility" entail "can avoid," the "can" in "can avoid" also expresses specific ability. In Frankfurt examples, however, agents like Laleh in stage 2 lack the specific ability to refrain from doing what they cannot avoid doing. In addition, should the objector think that Frankfurt cases show that obligation does not require specific ability but some other sort of ability, expose what this other variety of ability is – surely, it can't be general ability – and then I will be happy to attempt to assess what is on offer.

To summarize, the offender's relevant action in pertinent Frankfurt cases (for example, Laleh's ridiculing Arman in *Forgiveness*) is not impermissible for her (given *OIC* and *Equivalence*). You might, next, reason in these steps:

Argument 1

(1W) Since neither offender nor victim act any differently in a Frankfurt case than they do in a non-Frankfurt version of the case with no fail-safe mechanism (essentially stage 1), if forgiveness is not problematic in the non-Frankfurt case, it should not be problematic in the Frankfurt case either. (2W) As the victim still legitimately forgives in stage 2, forgiveness does not presuppose impermissibility. (3W) Given the choice between the objective impermissibility requirement and the subjective belief-in-impermissibility requirement, the latter is to be preferred.

When we forgive, however, the view that the offender has done something that is negatively valenced is not withdrawn. Something of a negative nature remains constant. If it is not wrongdoing that persists, one may recommend instead that what remains constant is the victim's belief that the offender has done wrong and that the offender is either blameworthy or the victim continues to believe that the offender is blameworthy.

Forgiveness and Blameworthiness

Turning now to the blameworthiness contrast pair, which of its members is more plausible: necessarily, to be forgiven for something (i) the offender is blameworthy for this thing and the victim believes that the offender is blameworthy for it, or (ii) the offender takes herself to be blameworthy for this thing and the victim believes that the offender takes herself to be blameworthy? An argument against the latter, which parallels the argument in the previous paragraph against the objective wrongness view, is problematic. To explain, appealing to *Forgiveness*, ponder the following against what may be labelled the "objective blameworthiness view":

Argument 2

(1B1) Since neither offender nor victim act any differently in a Frankfurt case than they do in a non-Frankfurt version of the case, if forgiveness is not problematic in the non-Frankfurt case (recall, in stage 1, the agent *is* blameworthy), it should not be problematic in the Frankfurt case either. (2B1) As the victim still legitimately forgives in stage 2, forgiveness does not presuppose blameworthiness. (3B1) Given the choice between the objective blameworthiness requirement and the subjective belief-in-blameworthiness requirement, the latter is to be preferred.

Frankfurt partisans should reject (2B1), insisting that Laleh remains blameworthy in stage 2 even though she could not have refrained from ridiculing Arman. After all, Frankfurt examples have been commissioned to undermine, among other versions of *PAP-R*, the principle (*PAP-Blame*) that you are blameworthy for doing something only if you could have done otherwise. How do we proceed in the face of this difficulty?

Let's look more closely at an alleged conceptual connection between blameworthiness and impermissibility, the principle that you cannot be blameworthy for something unless it is impermissible for you:

> *Blameworthiness Requires Impermissibility* (*BRI*): You are blameworthy for *A*-ing only if it is impermissible for you to do *A*.[1]

Should Frankfurt examples like *Forgiveness* at least initially sway you against the view that blameworthiness requires avoidability, you have reason to reject *BRI*. Recall, from *OIC* ("ought" implies "can") and *Equivalence* ("ought not"

1. E.g., Smith, 1991, 271; Widerker, 1991, 223; Fields, 1994; Copp, 1997; Copp, 2003, 286-287; Fischer, 2006, 218; Arpaly, 2006, 91n3; Campbell, 2011, 33-34; Nelkin, 2011, 100, 104; Franklin, 2018, 35; Brink, 2021, 117.

is equivalent to "wrong"), we infer *IAvoid* (if it is impermissible for you to do *A*, then you can avoid doing *A*). In addition, *OIC*, *Equivalence*, and *BRI* jointly entail *PAP-Blame*: If you are blameworthy for doing *A*, it is impermissible for you to do *A* (*BRI*). If it is impermissible for you to do *A*, you can refrain from doing *A* (*IAvoid*). Hence, if you are blameworthy for doing *A*, you can refrain from doing *A* (*PAP-Blame*).[1] Provided you are convinced that Frankfurt examples do, indeed, undermine *PAP-Blame*, since *PAP-Blame* follows from the conjunction of *OIC*, *Equivalence*, and *BRI*, you should conclude that one or more of these principles is false. If you accept *OIC* and *Equivalence*, you should reject *BRI*.[2]

There are additional putative reasons to question *BRI*. Some have proposed that *BRI* falls to counterexample (See, e.g., Zimmerman, 1997). In addition, suberogatory acts are, roughly, morally permissible acts that are especially disvaluable in some way and for the doing of which you are blameworthy. One may attempt to undermine *BRI* by defending the possibility that some actions are suberogatory for agents.[3] Yet others may reject *BRI* by defending an account of blameworthiness that invalidates *BRI*.

Several candidates have been proposed as replacements for *BRI*.[4] I assume that the strongest of these ties blameworthiness to belief in impermissibility:

> *Blameworthiness Requires Belief In Impermissibility* (*BRBI*): You are blameworthy for *A*-ing only if you believe that it is impermissible for you to do *A*.[5]

1. E.g., Widerker, 1991; Haji, 1994; Nelkin, 2011, 100-101.
2. Fischer (2006, 217-222) proposes that Frankfurt examples impugn *OIC*. (see Haji, 2019, 39-42 for a reply.) If you reject *OIC*, the preceding and subsequent reasoning that appeals to it will not convince you.
3. Indeed, I think that you can be blameworthy for doing *A* even if *A* is not suberogatory for you but you believe that *A* is suberogatory for you. On the suberogatory, see, e.g., Mellema 1991; Driver, 1992; McNamara, 1996; Haji, 2016, pp. 88-91.
4. Here are some: (1) You are blameworthy for *A*-ing only if you have a sufficient reason not to *A* and it is reasonable to expect you to know this. See, e.g., Fischer, 2006, 221. (2) You are blameworthy for *A*-ing only if it is bad for you to do *A*, or, the state of affairs, *your doing A*, is bad (or, maybe, *A* is bad). (A belief variant of (2) says: You are blameworthy for *A*-ing only if you believe that the state of affairs, *your doing A*, is bad or you believe that *A* is bad, or you do *A* in light of the belief that *A* is bad.) (3) You are blameworthy for *A*-ing only if you do *A* for morally bad or questionable reasons. E.g., Hanser, 2005, 454, 456; Nelkin, 2011, 105; Pereboom, 2014, 143. (A belief variant says: You are blameworthy for *A*-ing only if you do *A* for what you take to be morally bad or questionable reasons.) See Haji, 2023, 31-37, for critical discussion of these principles.
5. E.g., Moore 1912, 101; Parfit 1984, 25; Thomson 1991, 295; Haji, 2016, 165-190; Zimmerman, 1997; Hanser, 2005, 454, 456. Pereboom rejects *BRI* (2014, 142).

For simplicity, I ignore variations of *BRBI* that qualify it in certain ways, for example, that your belief that it is impermissible for you to do *A* is nonculpable or justified because these variations don't bear on the subsequent discussion.

Remaining neutral on whether blameworthiness requires impermissibility (*BRI*) or mere belief in impermissibility (*BRBI*), I advance the following "bridge principle" that associates blameworthiness with impermissibility.

> *Blameworthiness/Impermissibility*: You are blameworthy for *A*-ing only if either it is impermissible for you to do *A* (*BRI*) or you believe that it is impermissible for you to do *A* (*BRBI*).[1]

Returning to whether forgiveness presupposes blameworthiness (the objective blameworthiness view) or belief in blameworthiness (the subjective blameworthiness alternative), assume that one of *BRI* or *BRBI* is true. Suppose *BRI* is true. Then one might advance this argument against the objective blameworthiness option:

Argument 3

(1B2) Since neither offender nor victim act any differently in a Frankfurt case than they do in a non-Frankfurt version of the case, if forgiveness is not problematic in the non-Frankfurt case, it should not be problematic in the Frankfurt case either. (2B2) As the victim still legitimately forgives in stage 2, forgiveness does not presuppose blameworthiness. Expanding on the rationale for (2B2), since the offender's act is *not* impermissible for her in stage 2 (as previously explained), and *BRI* (the principle that blameworthiness requires impermissibility) is assumed to be true, the offender is not blameworthy for what she does in stage 2. But if the victim still forgives in stage 2, forgiveness does not presuppose blameworthiness. (3B2) Given the choice between the objective blameworthiness requirement and the subjective belief-in-blameworthiness requirement, the latter wins.

Supposing *BRI* is true, it seems that (2B2) is no longer controversial provided that the offender's action is not impermissible for her in stage 2. However, there is still a left-over worry with *Argument 3*. Although *OIC* and *Equivalence* are relatively unproblematic, *BRI* is far more contentious. Indeed, as previously cautioned, Frankfurt cases themselves call *BRI* into question. (3B2) may, nevertheless, still be accepted because the alternative to *BRI* is

1. Maybe there is some complicated principle that conceptually ties blameworthiness to, for instance, all three of badness, bad reasons, and impermissibility. I assume any such principle (if there is one) entails one or both of *Blameworthiness/Impermissibility*'s disjuncts.

BRBI. If blameworthiness requires belief in impermissibility (*BRBI*) and not impermissibility *per se* (*BRI*), and if we grant that Arman does forgive Laleh in stage 2 and there is nothing suspect about this instance of forgiveness, then forgiveness is tied to belief in blameworthiness and not to blameworthiness *per se*. (Take it as given that Arman believes that Laleh is blameworthy for ridiculing him.)

Collecting results, grant that Arman *does* forgive Laleh in Frankfurt case *Forgiveness*. If *BRI* (blameworthiness requires impermissibility) is true, this should give us reason to accept the subjective belief-in-blameworthiness requirement on forgiveness as *Argument 3* shows. If *BRI* is false, and its alternative *BRBI* (blameworthiness requires belief in impermissibility) is true, this again gives us reason to accept the belief-in-blameworthiness requirement on forgiveness. Unless one has *antecedently* made up one's mind about the truth of both the objective impermissibility requirement and the objective blameworthiness requirement on forgiveness, the competing claim that their subjective rivals should replace both these putative requirements is *prima facie* reasonable. In addition, suppose we also accept principles *OIC* and *Equivalence*. Then the contender that forgiveness requires belief in impermissibility and not impermissibility *per se* is plausible as is the candidate that forgiveness requires belief in blameworthiness rather than blameworthiness *per se*.

Forgiveness and Determinism

The stage is now set to draw some conclusions about whether determinism imperils forgiveness. Take as unproblematic that determinism undermines freedom to do otherwise (See, e.g., Wiggins, 1973; van Inwagen, 1983; Ginet, 2003). Bridge principle *Blameworthiness / Impermissibility* gives us two alternatives regarding blameworthiness and impermissibility: either blameworthiness requires impermissibility (*BRI*) or belief in impermissibility (*BRBI*). Suppose *BRI* is true. Since *OIC* and *Equivalence* entail that nothing can be impermissible for one unless one could have done otherwise, and determinism precludes freedom to do otherwise, *prima facie*, nothing can be impermissible for anyone in a deterministic world.[1] Hence, if blameworthiness requires impermissibility, no one can be blameworthy for anything in a deterministic world.

Assume, next, that forgiveness requires both impermissibility and

1. I say "prima facie" because compatibilists may well wish to defend the view that that "can" of obligation is some weak compatibilist "can" and not some relatively strong incompatibilist "can." Compatibilists of this bent can accept the view that nothing can be impermissible or obligatory for you unless you could have done otherwise, and affirm that determinism does not preclude you from doing otherwise, given an apt compatibilist analysis of *one's being able to do otherwise*.

blameworthiness. To be forgiven, the offender must have done something morally wrong and for which she is to blame. Since deterministic worlds are shorn of impermissibility and blameworthiness, there is no fitting or apt forgiveness in such worlds. You may initially think that this is nothing momentous because it is the subjective member of each contrast pair that is true: necessarily, to be forgiven, the victim and offender believe that the offender has done wrong and believe that the offender takes herself to be blameworthy for this wrongdoing. Nothing precludes agents from having false beliefs in a deterministic world: although such a world is devoid of blameworthiness and wrongness, agents in this world may act on the basis of the false belief that they are doing wrong and falsely believe that they are blameworthy for their wrongdoing, and others may falsely believe that these agents have done wrong (or take themselves to have done wrong) and are to blame (or believe that they are to blame) for their wrongdoings. Hence, if the subjective members of either contrast pair are true, determinism does not preclude forgiveness.

However, even if the subjective members of the contrast pairs are to be preferred to their objective members, there is a bitter pill to swallow regarding forgiveness in deterministic worlds: there is a transparent sense in which judgments of blameworthiness, impermissibility, and forgiveness would be irrational in these worlds insofar as these judgments would be predicated on false beliefs, for example, the false belief that acts can be wrong for agents in these worlds. A practical corollary of this sobering conclusion is that if agents in deterministic worlds come to the realization, perhaps over time, that nothing they do is impermissible for them (or that everything they do is amoral for them), and that to be blameworthy for something in their deterministic world, they act on the false belief that they do wrong, no agent would be blameworthy for anything should agents give up these false beliefs. It would seem that knowing that nothing they do is impermissible for them, these agents (again, perhaps over time) would fail to act on the basis of the belief that they are doing wrong, and would fail to believe that agents in this world act on the basis of the belief that they are doing wrong. Hence, even if blameworthiness requires belief in impermissibility, it would be unrealistic to assume that these agents would ever be blameworthy for anything and that these agents would believe that anyone is blameworthy for anything. It follows that even if the subjective members of the contrast pairs are true, forgiveness would fade away in deterministic worlds.

Forgiveness and Artificial Agency

The wrongness contrast pair says that, necessarily, to be forgiven for something (i) it's wrong for the offender to do this, or (ii) the offender believes

that it's wrong for her to do this. The blameworthiness pair affirms that, necessarily, to be forgiven for something (i) the offender is blameworthy for this thing, or (ii) the offender believes that she's blameworthy for this thing. I have argued in favour of the second members – the belief or subjective members – of each pair. Regarding whether AI agents may be apt candidates of forgiveness, I remain agnostic about which member of each pair is true to guide the discussion.[1] The relevant issues to be tackled for what ensues are whether AI agents can do wrong or have relevant beliefs about wrongdoing, beliefs such as it is morally impermissible for me, R2D2, to make this decision, and whether AI agents can be blameworthy for at least some of their decisions or have the pertinent beliefs about being blameworthy. I assume (dogmatically) that AI agents are at least minimal agents – they can perform intentional actions, notably, mental actions of decision-making.

AI Agents and Wrongdoing

Sharif, an adorable Old English Sheepdog, performs many intentional actions (I take this to be uncontroversial) but does he ever do wrong? Compare: does a one-year-old human child ever do wrong? Does an adult chimpanzee ever do wrong? The challenging well-known dispute here concerns what may be dubbed "moral normative agency": what conditions must a minimal agent satisfy for at least some of her actions to instantiate the morally deontic properties of rightness, wrongness, or obligatoriness? The strategy I deploy to make some headway starts with the issue of "thick normative agency," which concerns relatively strong agency requirements of moral responsibility.

One such requirement is that the candidate be capable of intentional deliberative action. Such action, in turn, requires some psychological basis for evaluative reasoning. To be responsible, for example, one must, at least on occasion, be able to assess which line of conduct is best from one's own perspective of what is best. An agent's deliberations involve an appraisal of reasons for or against action. They issue in a practical judgment, which in turn gives rise to a decision or intention.

Drawing on previous work, I propose that an agent requires an *evaluative scheme* (or constituents of such a scheme) to entertain such reasons and to appraise them (see, e.g., Haji & Cuypers, 2008, ch. 3). An evaluative scheme has these constituents: (a) Normative standards the agent believes (though not necessarily consciously) ought to be invoked in assessing reasons for action, or beliefs about how she should go about making choices. Leave it open what

1. As previously noted, I leave it open that both members may be false.

these standards might be, or how much weight one gives to standards that, on occasion, deliver conflicting verdicts. To be a fitting candidate for *moral* responsibility, an agent's evaluative scheme's deliberative standards must include a set of *moral* principles or norms to ensure minimal moral competence: the agent must grasp the concepts of rightness, obligatoriness, or wrongness and be able to appraise, morally, various choices or actions in light of the moral norms that are elements of her evaluative scheme. There is no requirement that these moral appraisals be fully considered, free of error, or even conscious. Nor is there any requirement that these norms be evidentially based or justified in any strong sense of 'justification.' The agent may simply assimilate, without critical scrutiny, various norms of her religion or culture. (b) Included in the agent's evaluative scheme are desires, beliefs, or plans that express the agent's long-term ends or goals she deems worthwhile or valuable. For example, an agent may underscore her commitment to attempt to maximize overall happiness whenever she acts. (c) An evaluative scheme also has, as constituents, deliberative principles that the agent utilizes to arrive at practical judgments about what to do or how to act. For instance, the agent may endorse a principle of this sort: when self-interest conflicts with the demands of peer pressure, always side with self-interest. Or one may believe that the best way to maximize overall happiness is to rely on rules of thumb, such as, "keep your promises," "don't cheat," "don't steal," and so forth. (d) Lastly, an evaluative scheme incorporates motivation both to act on the normative standards specified in (a) and to pursue one's goals of the sort described in (b) at least partly on the basis of engaging the deliberative principles outlined in (c).

One may opt to refine (a), (b), (c), and (d). One might, for instance, wonder whether satisfaction of any of these conditions entails that the agent is self-conscious. Here, I rest (uneasily) with what has been documented about (a), (b), (c), and (d), fully recognizing that far more needs to be said to appreciate the complex notion of *being a normative agent*.

It *suffices* for an individual's being a normative moral agent at a certain time that the individual have at that time (i) an evaluative scheme with the requisite moral elements – the agent is minimally morally competent; (ii) deliberative skills and capacities; for example, the agent has the capacity to apply the normative standards that are elements of her evaluative scheme to assess reasons; and (iii) executive capacities – the agent is able to act on at least *some* of her intentions, decisions, or choices; she has the capacity to translate some of her intentions, decisions, or choices into action. An individual, like a toddler, who fails to have deliberative or executive capacities will be able to exert much less control, if any, over its actions than an

individual who does have such capacities. Interpret condition (ii) to entail that the agent can engage in *genuine* deliberation; her deliberative activities must meet the threshold of rationality below which such activities fail to count as bona fide *deliberation*. Persons suffering from certain severe psychological ailments may not be capable of genuine deliberation.

Normal, mentally healthy, adult human beings are normative moral agents. Sheep dogs, like Sharif, and human babies are not. Presumably, Sharif is not minimally morally competent. Not having moral standards, Sharif cannot apply any such standards to evaluating reasons for action even if we assume that he does undertake crude evaluations of his reasons. A plausible proposal is that an entity or agent can perform actions that are right, wrong, or obligatory only if it is a normative moral agent, and one is a normative moral agent only if one is minimally morally competent. On this moral competence model, to be a normative moral agent, one must have a basic grasp of the concepts of rightness, wrongness, and obligatoriness, and one must be able to use these concepts to appraise various things.

Although I am attracted to the moral competence model of deontic moral agency, I suspect that its requirements are too stringent: things that fail to be minimally morally competent may, it seems, perform actions that can be right, wrong, or obligatory. Think, for instance, of hypothetical agents who are devoid of moral concepts, do perform intentional actions, and undertake aesthetic or prudential evaluations of various things like policies, courses of action, paintings, architecture, and the like. Such beings would be normative agents of a particular variety: though normatively competent, their normative competence would not include moral competence. These beings would have an evaluative scheme with appropriate aesthetic and prudential elements but not moral elements, deliberative skills and capacities, and executive capacities. Intuitively, such beings may well perform actions that are morally wrong, although they wouldn't know that these were morally wrong, or they wouldn't perform them in light of the belief that they were doing moral wrong. With this sort of example in mind, one might offer, in contrast to the moral competence model, a normative competence model according to which an entity can perform actions that are right, wrong, or obligatory only if it is a *bona fide* normative agent of some sort; it must have an evaluative scheme that includes normative standards for evaluation, deliberative skills and capacities, and executive capacities. Such an entity would be able to undertake some sort of normative appraisal – one that needn't be moral – of things like policies, actions, and so forth.

It may be objected that whereas this relatively weak model gets the case of the hypothetical beings right, one of its drawbacks is that it rules that a tiger's act of, say, killing a human child, is wrong. But this is controversial. It is

debatable, for instance, whether tigers have anything like an evaluative scheme, even one that need not include moral standards of evaluation.

I shall not decide between the two models though, of the two, I'm drawn to the weaker second. Reverting to AI agents, and focusing, first, on whether they can do wrong, or act on the basis of the belief that they are doing wrong, one would need to ascertain whether AI agents possessing deliberative and executive capacities and equipped with evaluative schemes that include appropriate moral elements can be constructed. Speculating, I see no reason to deny that it is possible to fabricate such agents.

AI Agents and Blameworthiness

Can AI agents be built that are the sort of agent capable of commissions or omissions for which they can be morally blameworthy or, more generally, morally responsible? Could they or their relevant behavior satisfy the conditions of moral responsibility? A non-exhaustive list of these conditions includes: (1) A control or freedom condition. (2) An agency condition. (3) Many have proposed an ownership condition: roughly, the springs of action that issue in one's decisions must be "truly one's own" and not, for instance, the product of covert manipulation of which one is unaware and cannot thwart (see, e.g., Haji & Cuypers, 2008, chs. 2-3; Kane, 1996, ch. 5). (4) An epistemic requirement perhaps along the lines that the agent believes that she is doing wrong. (5) What may be dubbed a "moral requirement," alleged examples of which are the blameworthiness-requires-impermissibility condition or blameworthiness-requires-belief-in-impermissibility condition. I offer this list as an invitation for further discussion on whether anything prevents development of AI agents who meet these conditions. This task is monumental. To expose just a few of the challenges, there has been long-standing debate raging over each of these conditions. For instance, regarding control, does responsibility require reasons-responsiveness, being able to conform one's actions to what one values or to desires with which one identifies, or being able to do something other than what one does, given the same past and the laws of nature? As for the moral requirement, does blameworthiness – the negative mode of responsibility – require wrongdoing or merely belief in wrongdoing? Clarity on these conditions is simply the initial step. One would then have to ascertain whether it's possible for AI agents to satisfy these conditions. I very gladly leave these (and the requisite attendant) tasks to others.

▣ Conflict of Interest

The author declares no competing interests.

References

Allais, L. (2008). Wiping the slate clean: The heart of forgiveness. *Philosophy & Public Affairs, 36*(1), 33–68. https://doi.org/10.1111/j.1088-4963.2008.00123.x

Allais, L. (2014). Freedom and forgiveness. In D. Shoemaker & N. Tognazzini (Eds.), *Oxford studies in agency and responsibility* (Vol. 2, pp. 33–63). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198722120.003.0004

Arpaly, N. (2006). *Merit, meaning, and human bondage: An essay on free will*. Princeton University Press.

Belzer, M. (1998). Deontic logic. In *Routledge encyclopedia of philosophy*. Taylor & Francis. https://www.rep.routledge.com/articles/thematic/deontic-logic/v-1

Brink, D. (2021). *Fair opportunity, responsibility, and excuse*. Oxford University Press.

Campbell, J. K. (2011). *Free will*. Polity Press.

Clarke, R. (2015). Abilities to act. *Philosophy Compass, 10*(12), 893–904. https://doi.org/10.1111/phc3.12282

Copp, D. (1997). Defending the principle of alternate possibilities: Blameworthiness and moral responsibility. *Noûs, 31*(4), 441–456. https://doi.org/10.1111/0029-4624.00055

Copp, D. (2003). "Ought" implies "can": Blameworthiness and the principle of alternative possibilities. In D. Widerker & M. McKenna (Eds.), *Moral responsibility and alternative possibilities: Essays on the importance of alternative possibilities* (pp. 265–299). Ashgate.

Driver, J. (1992). The suberogatory. *Australasian Journal of Philosophy, 70*(3), 286–295. https://doi.org/10.1080/00048409212345181

Feldman, F. (1986). *Doing the best we can*. D. Reidel.

Fields, L. (1994). Moral beliefs and blameworthiness. *Philosophy, 69*(270), 397–415. https://doi.org/10.1017/S0031819100047239

Fischer, J. M. (2006). *My way: Essays on moral responsibility*. Oxford University Press.

Frankfurt, H. G. (1969). Alternate possibilities and moral responsibility. *Journal of Philosophy, 66*(23), 829–839. https://doi.org/10.2307/2023833

Frankfurt, H. G. (1999). On caring. In *Necessity, volition, and love* (pp. 155–180). Cambridge University Press.

Franklin, C. (2018). *A minimal libertarianism: Free will and the promise of reduction*. Oxford University Press.

Ginet, C. (2003). Libertarianism. In M. J. Loux & D. W. Zimmerman (Eds.), *The Oxford handbook of metaphysics* (pp. 252–271). Oxford University Press.

Haji, I. (1994). Changing obligations and immutable blameworthiness. *Theoria, 60*(1), 48–62. https://doi.org/10.1111/j.1755-2567.1994.tb00877.x

Haji, I. (2002). *Deontic morality and control*. Cambridge University Press.

Haji, I. (2006). Frankfurt examples, obligation, and responsibility. *Journal of Ethics, 10*(3), 255–281. https://doi.org/10.1007/s10892-005-5781-1

Haji, I. (2016). *Luck's mischief*. Oxford University Press.

Haji, I. (2019). *The obligation dilemma*. Oxford University Press.

Haji, I. (2023). *Obligation and responsibility*. Oxford University Press.

Haji, I., & Cuypers, S. (2008). *Moral responsibility, authenticity, and education*. Routledge.

Hanser, M. (2005). Permissibility and practical inference. *Ethics, 115*(3), 443–470. https://doi.org/10.1086/428457

Hebert, R. (2016). *Teleologism full stop: A general theory of ability, agency, know-how, obligation, and justification* [Doctoral dissertation, University of Calgary].

Hieronymi, P. (2001). Articulating an uncompromising forgiveness. *Philosophy and Phenomenological Research, 62*(3), 529–555. https://doi.org/10.1111/j.1933-1592.2001.tb00073.x

Hilpinen, R., & McNamara, P. (2013). Deontic logic: A historical survey and introduction. In D. M. Gabbay, J. F. Horty, X. Parent, R. van der Meyden, & L. van der Torre (Eds.), *Handbook of deontic logic and normative systems* (pp. 3–36). College Publications.

Holmgren, M. R. (1993). Forgiveness and the intrinsic value of persons. *American Philosophical Quarterly, 30*(4), 341–351.

Holmgren, M. R. (2012). *Forgiveness and retribution: Responding to wrongdoing*. Cambridge University Press.

Hughes, P. M. (1993). What is involved in forgiving? *Journal of Value Inquiry, 27*(3), 331–340. https://doi.org/10.1007/BF01087682

Kane, R. (1996). *The significance of free will*. Oxford University Press.

Littlejohn, C. (2012). Does "ought" still imply "can"? *Philosophia, 40*(4), 821–828. https://doi.org/10.1007/s11406-012-9366-5

McNamara, P. (1996). Making room for going beyond the call. *Mind, 105*(420), 415–450.

McNamara, P. (2019). Deontic logic. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Summer 2019 ed.). https://plato.stanford.edu/archives/sum2019/entries/logic-deontic/ (Original work published 2006)

Mellema, G. (1991). *Beyond the call of duty: Supererogation, obligation, and offence*. State University of New York Press.

Moore, G. E. (1912). *Ethics* (W. H. Shaw, Ed.). Oxford University Press.

Murphy, J. G., & Hampton, J. (1988). *Forgiveness and mercy*. Cambridge University Press.

Nelkin, D. K. (2013). Freedom and forgiveness. In I. Haji & J. Caouette (Eds.), *Free will and moral responsibility* (pp. 165–188). Cambridge Scholars Publishing.

Nelkin, D. K. (2011). *Making sense of freedom and responsibility*. Oxford University Press.

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

Pereboom, D. (2001). *Living without free will*. Cambridge University Press.

Pereboom, D. (2014). *Free will, agency, and meaning in life*. Oxford University Press.

Pereboom, D. (2021). *Wrongdoing and the moral emotions*. Oxford University Press.

Rönnedal, D. (2009). *An introduction to deontic logic*. CreateSpace Independent Publishing Platform.

Smith, H. (1991). Varieties of moral worth and moral credit. *Ethics, 101*(2), 279–303. https://doi.org/10.1086/293289

Thomson, J. J. (1991). Self-defense. *Philosophy & Public Affairs, 20*(4), 283–310.

van Inwagen, P. (1983). *An essay on free will*. Clarendon Press.

Vihvelin, K. (2004). Free will demystified: A dispositional account. *Philosophical Topics, 32*(1–2), 427–450. https://doi.org/10.5840/philtopics2004321/211

Vihvelin, K. (2013). *Causes, laws, and free will*. Oxford University Press.

Vranas, P. (2007). I ought, therefore I can. *Philosophical Studies, 136*(2), 167–216. https://doi/org/ 10.1007/s11098-007-9071-6

Warmke, B. (2016). The normative significance of forgiveness. *Australasian Journal of Philosophy, 94*(1), 1–17. https://doi.org/10.1080/00048402.2015.1126850

Warmke, B., & McKenna, M. (2013). Moral responsibility, forgiveness, and conversation. In I. Haji & J. Caouette (Eds.), *Free will and moral responsibility* (pp. 189–212). Cambridge Scholars Publishing.

Widerker, D. (1991). Frankfurt on "ought implies can" and alternative possibilities. *Analysis, 51*(4), 222–224. https://doi.org/10.1093/analys/51.4.222

Wiggins, D. (1973). Towards a reasonable libertarianism. In T. Honderich (Ed.), *Essays on freedom of action* (pp. 31–61). Routledge & Kegan Paul.

Zimmerman, M. J. (1996). *The concept of moral obligation*. Cambridge University Press.

Zimmerman, M. J. (1997). A plea for excuses. *American Philosophical Quarterly, 34*(2), 229–243.

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

The Importance of Being Conscious in the Debate over AI

Charles Taliaferro

Professor Emeritus of Philosophy, St. Olaf College, Northfield, Minnesota, USA

taliafer@stolaf.edu

Abstract

This paper defends the evident reality of consciousness, including the existence of conscious, enduring selves, against recent objections raised by Helen Yetter-Chappell and Anil Seth. Yetter-Chappell advances a non-theistic idealism that echoes Hume's bundle theory, denying a substantial self behind experience, while Seth characterizes the self as a controlled hallucination generated by predictive brain processes. I argue that both accounts mischaracterize the common-sense phenomenology of self-awareness: when one reads, thinks, or feels pain, one is directly aware of oneself as a substantial, enduring subject – not of a "lurking" entity behind experience. I further contend that Seth's rejection of appearances is epistemically precarious and that his analogy between consciousness and heat fails to explain the subjective sensation of hotness. The paper then turns to the debate over artificial intelligence. If consciousness is bracketed and intelligence is analyzed behavioristically or functionally, AI may indeed equal or surpass human intelligence; however, this is profoundly different from being intelligent in the sense of conscious self-awareness. Through a critical engagement with Daniel Dennett's thought experiment of the unconscious robot "Zeke," I show that consciousness is both evident and indispensable to what we value in friendship, love, work, and moral life. The paper concludes that while behaviorism and functionalism may grant intelligence to AI, they cannot capture what it means to be a conscious, self-aware being – a reality of paramount importance for how we understand ourselves, other creatures, and God.

Original Research

Keywords:

consciousness, intelligence, Yetter-Chappell, Anil Seth, artificial intelligence, theism.

Received: 2026/02/05; **Received in revised form:** 2026/04/12; **Accepted:** 2026/05/27; **Published online:** 2026/06/10

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), 21-38. https://doi.org/10.22091/jptr.2026.17505.3695

Copyright © The Author

Introduction

Consciousness plays an essential role in world religions. In some forms of Hinduism, Brahman is pure, infinite consciousness. Buddha-consciousness is a central notion in most forms of Buddhism. In the Abrahamic faiths, God's creation and sustaining of the cosmos is a conscious divine act, and God's omniscience consists in God's awareness of all dimensions of the created order. God's justice and mercy, God's love and providential care are all within the scope of God's boundless conscious life. These traditions also stress the importance of our consciously responding in intentional acts of prayer, worship, repentance, reverence, loving and caring for the creation. We are to live life in light of God's presence (coram Deo), in the faith and hope for life after this life (Taliaferro, 2024).

In the book *Consciousness and the Mind of God,* I argued that the credibility of theism is closely tied to the credibility of believing in the reality of human consciousness (Taliaferro, 1994). If certain forms of eliminative materialism give us compelling reasons to doubt the reality of consciousness in humans and some nonhuman animals, then God's consciousness should also come under suspicion.. Something similar is true in the current debate over artificial intelligence.

If we bracket the reality of consciousness and assume, for example, a behaviorist or functional analysis of intelligence, then AI has already equalled or surpassed human intelligence, or will soon do so. AI has surpassed us in solving mathematical problems, composing texts, creating songs, practicing surgery, offering therapy,planning , and even conducting wars with the use of mechanical flying robots or drones, and so on.

In this paper, I submit there are compelling reasons to recognize the reality of ourselves as self-aware conscious subjects who think, reason, have sensation, experienced emotions, agency, memory, and so on over time. I then contend that the reality of consciousness plays a vital role in our assessment of AI.

Conscious Selves

Along with many other philosophers, including David Lund, Roderick Chisholm, Stewart Goetz, J. P. Moreland, and others, I contend that we are immediately aware of ourselves enduring over time with an array of sensations, thoughts, feelings, and possessing the power of agency as physically embodied beings (Loose, Menuge & Moreland, 2018). In order not to replicate what I have published in other books, in this paper I will defend the

reality of conscious selves against objections raised by two philosophers who present recent challenges to the position I hold: Helen Yetter-Chappell, in *The view from Everywhere: Realist Idealism without God,* and Anil Seth, in *Being You: A New Science of Consciousness.*

Yetter-Chappell articulates and defends the coherence and plausibility of a form of idealism that does not require a theistic basis. Whatever its merits (and there are many), I focus on her view of the conscious self.

In several places, Yetter-Chappell supports David Hume's bundle theory of the self (Yetter-Chappell, 2025, pp. 17, 65). Famously, Hume claimed he never perceived himself, but only perceptions of pleasure or pain, love or hatred, heat or cold, and so on (as cited in Yetter-Chappell, 2025, p. 17). She echoes Hume's claim in her own experience:

> Like [Mark] Johnson (and Hume), I fail to find a self lurking about my mental acts, as something over and above what I experience. I find the computer before me, my hands typing, the thoughts running through my head, the sound of a bird, the ache in my left calf. I do not find the computer in its entirety, but a particular perspective on the computer. All these experiences are unified into the total experience of me at a moment (Yetter-Chappell, 2025, p. 173).

She distinguishes three positions in terms of the self and experience, which she calls the robust mind, the thin mind, and no mind. The first refers to the mind or self as a substantial reality, an individual subject. The second is described as an experiencer (some kind of subject?), but not something that adds to an ontology of experience. The third is the denial that there is any kind of subject.

Claims like Hume's – and, by extension, Yetter-Chappell's – have bewildered many (but not all) philosophers. Arguably, any of Hume's claims about perception (or emotion or sense) seem inextricably to be examples of self-awareness. Who can experience pain without the person feeling herself in pain? What sense can be made of a person claiming to feel hot without her being committed to the view that she herself is feeling hot? Yetter-Chappell's terminology seems skewed. Those who claim that we are aware of ourselves as substantial subjects enduring over time would make no sense of the idea of a lurking self behind Yetter-Chappell's experiences. In her description of her own experiences, who sees the computer if she does not? When she is typing, doesn't she have an awareness that it is not just her hands that are typing, but that she is the one typing? To read her account, cited above, don't you need to experience yourself (a substantial individual person who endures over time) as reading the whole text over time? The so-called "thin mind" seems (at least to

me) preposterous. How might you claim there is a person (a reader) who is reading the text, but there is no ontological significance to the claim there is more than the experience of reading (namely, there is a reader)?

My challenge to Yetter-Chappell is philosophical and based on what seems to me (and many others) common-sense phenomenology. But the neuropsychologist Oliver Sacks has proposed that some severe pathology can so disturb a person's inner life that something like a Humean predicament can seem real.

"For Hume, personal identity is a fiction – we do not exist, we are but a consecution of sensations, or perceptions. This is clearly not the case with a normal human being, because he owns his own perceptions. They are not a mere flux but his own, united by an abiding individuality or self. But what Hume describes may be precisely the case for a being as unstable as a super-Touretter, whose life is, to some extent, a consecution of random or convulsive perceptions and motions, a phantasmagoric fluttering with no center or sense. To this extent, he is a 'Humean' rather than a human being. This is the philosophical, even theological fate which lies in wait, if the ratio of impulse to self is too overwhelming. It has affinities to a 'Freudian' fate, which is also overwhelmed by impulse – but the Freudian fate has sense (albeit tragic), whereas a 'Humean' fate is meaningless and absurd" (Sacks, 1985, p. 146).

Let us now turn to Anil Seth. In *Being You,* he acknowledges that we appear to exist as conscious subjects but raises the question whether this is an illusion: "The notion that there is a single unique conscious self (a soul?) that persists over time may be grossly mistaken" (Seth, 2022, p. 4). In his view, we are "Beast machines," and being you and me emerges "from the way our brain predicts and controls the internal states of the body" (Seth, 2022, p. 6). Like Yetter-Chappell, Seth construes those of us who believe we have immediate self-awareness as positing a lurking self; he charges: "The self is not some immutable entity that lurks behind the shadows of the eyes" (Seth, 2022, p. 154). He rejects common-sense appearances: "the way things seem is often a poor guide to how they actually are" (Seth, 2022, p. 17).

Seth's rejection of common-sense appearances is also reflected in his account of our experience of colors (red is his specific example) and his view that our perception of the world should be thought of as what he calls a controlled hallucination. He claims that what it is like to see red (or see something red) should be largely understood as an experience of what it is not (blue, green, etc.) (Seth, 2022, p. 53). He grants that there is a phenomenological subjective experience of red (Seth, 2022, p. 85), but it is part of a neurological hallucination: "The entirety of perceptual experience is a neuronal fantasy that remains yoked to the world through a continuous making

and remaking of perceptual best guesses, of controlled hallucinations" (Seth, 2022, p. 87).

Seth does not claim that we have established the identity of the mental with the physical, but he claims we are making progress. His example is the scientific discovery of heat being mean kinetic energy.

> The history of thermometry, and its impact on our understanding of heat, offers a vivid example of how the ability to make detailed, quantitative measurements, on a scale defined by fixed points, has the power to transform something mysterious into something comprehensible … If consciousness turns out to be like temperature – which is to say, if there is a single physical process that underlies and is identical to 'being conscious' – the payoff would be spectacular (Seth, 2022, p. 35).

Seth cites the fact that neurological damage can undermine our sense of being a unified self as evidence that there is no unified self. Under normal, healthy conditions, we experience ourselves as unified centers of consciousness (Seth, 2022, p. 54), but this is vulnerable to disintegration (Seth, 2022, pp. 149–150).

Seth ends his book offering some consolation about our death: Your death is the death of "a controlled hallucination" (Seth, 2022, p. 276). We will all pass into oblivion or nothingness, but this is "a return to the eternity that each of us at some time emerged from" (Seth, 2022, p. 276).

I offer five replies.

First, Seth's depiction of those of us who believe in our reliable, direct awareness of ourselves is mistaken. We believe that self-awareness is not an inference to a *lurking* entity behind experience. Instead, we contend that when you read this sentence, you experience yourself as a substantial individual reader. Seth describes those of us who believe in a conscious self as maintaining that the self is immutable (cited above). Not at all. An immutable (non-changing) self would be akin to a statue or frozen self, not a living, changing, conscious being. Instead, we maintain that selves experience and cause change.

Second, Seth's repudiation of trusting appearances is precarious epistemically; it can put one on a slippery slope. Arguably, Seth has to rely on an appearance principle when he claims that our perceptual life is a controlled hallucination. That is, Seth is committed to the view that (at least in his case), based on his scientific view of the brain, *it appears that our perceptual life is a controlled hallucination.* A repudiation of appearances needs to be qualified, as in Smith's claim *that when he has taken LSD*, *the apparent reliability of his*

emotions (feeling paranoia, for example) *is less reliable than when he is not using psychotropic drugs.* In my view, Seth still needs to accept some version of an appearance principle (e.g., if some state of affairs appears to be the case to a subject, she has some reason to believe the state of affairs is true). (For a defense of appearance epistemology, see Taliaferro & Reasoner, 2026.)

Third, on redness and controlled hallucination: unless we defy all experience, we should (in my view) acknowledge the reality of our experience of redness and hallucinations. While I reject the view that our normal, healthy perception of each other and the world is a hallucination, I grant that people sometimes (under the influence of LSD, for example) actually hallucinate. While the hallucination of a butterfly is not a butterfly, it is a visual image. Seth seems to leave us in limbo about the status of visual images and fields. He acknowledges that when a person sees red, there is no redness out there in the world, nor is there any redness in the brain (Seth, 2022, p. 86). Well, where is it? One promising option is to claim that conscious selves perceive the world by means of a visual field or representation of the world.

Fourth, Seth's example of heat does nothing to advance the claim that consciousness is a brain process. The analysis of heat in terms of molecular motion is a fine depiction of non-sentient heat, but it does nothing to account for the actual conscious sensation of hotness.

Fifth, the evidence that our self-awareness is subject to distortion, fragmentation, or utter destruction due to all kinds of harms is no evidence at all that there are no unified, conscious selves. Actually, the embedded, intuitive appearance that we are conscious selves can come to the fore when we contemplate the death of someone we love. Shortly before I started writing this paper, a very dear friend of mine for many years, Michael, whom I love dearly, died. Speaking for myself, I find no consolation in believing Michael never was a conscious self, but a controlled hallucination who (or that?) has irreversibly perished.

Actually, while Seth seems to foreclose the possibility that Michael or any of us might survive death, he does think we should take reports of near-death and out-of-the-body experiences (OBEs) seriously.

> When people report supernatural or otherwise bizarre experiences, like OBEs, we should take seriously their reports. They probably do have the experiences they say they have. People have had real out-of-the-body experiences for millennia, but this does not mean that immaterial selves or immutable souls have ever actually left any physical bodies. What these reports reveal is that first-person perspectives are put together in more complex, provisional, and

> precarious ways than we will ever have direct subjective access to (Seth, 2022, p. 159).

This is not the place for me to philosophically assess OBEs. But I suggest that what these reports might also reveal is the enduring, ostensibly sensible idea that we are not controlled hallucinations but conscious selves who may not share the same fate as our bodies at death. I suggest that taking OBE accounts "seriously," should lead us to accept that persons claim to have direct subjective awareness of themselves some distance from their bodies, even if we contend that such experiences are caused by nero-chemical processes. Moreover, I have yet to come across a report of an OBE in which someone claimed they have experienced themselves as "immaterial selves or immutable souls." The reports available in the literature are far simpler, e.g. I witnessed physicians operating on my body from a distance (see Gary Hbermas 2023, 323-356).

AI withoutConsciousness

I have conceded at the outset that if we deny the reality of consciousness, and settle for some form of illusionism (there are no conscious selves, or the conscious self is an illusion or hallucination), then the status of AI is philosophically formidable. If some machines can outperform humans at most, if not all, tasks, shouldn't we treat them as persons or person-like creations, perhaps entitled to human rights and duties?

To develop this argument, let us consider a proposal advanced by Daniel Dennett.

Dennett asks us to imagine a time (perhaps sooner than we think) when scientists and philosophers will think that the belief in consciousness is akin to believing in geo-centricity, the idea that the sun rotates around an immovable earth. He envisions "a day when philosophers and scientists and lay-people will chuckle over the fossil traces of our earlier bafflement about consciousness. 'It still seems [they will say] as if these mechanistic theories of consciousness leave something out, but of course that's an illusion. They do in fact explain everything about consciousness that needs explanation'" (Dennett, 2001, p. 37).

To enhance his position, he asks us to imagine a robot he calls Zeke. Zeke is an AI who has all the (apparent) cognitive faculties we have; she has reasoning, sensory perception, memory, the power to move, speak, and hear, but all of these (as well as Zeke herself) are not conscious or, in Dennett's terminology, unconscious. While unconscious, Zeke has sensory awareness of

all her states. Dennett then imagines the following conversation with Zeke:

> "Zeke, do you like me?" /
> "Of course I do." /
> "Did you mind me asking?" /
> "Well, yes, it was almost insulting." /
> "How do you know?" /
> "I just recall feeling a bit annoyed or threatened." (Dennett, 2013, p. 291)

Dennett creates a scenario in which it seems impossible to tell whether Zeke is a conscious self or an unconscious machine. He asks us why we would care whether Zeke is conscious or not: "Or, more personably, why would you care whether you are, or became, a zimbo [a non-conscious or unconscious being]? In fact, you'd never know" (Dennett, 2013, p. 292).

If there is no such thing as consciousness, as a real state of self-awareness and awareness of love and the world, Dennett has a point. But I suggest the thought experiment can bring to light the evident reality and value of consciousness.

Consider three points.

First, beginning with a minor point, Dennett's analogy with geo-centricity can be misleading. He seems to assimilate those of us who persist in believing in conscious selves to those who persist in believing in a stationary earth with an orbiting sun. Dennett seems to be saying that while it appears that the earth is stationary, it is not. What is misleading here is that motion is measured in frameworks. In my current environment, a coffee house in North America, the *place on earth I am working is stationary.* This is not *a mere appearance,* something to outgrow with basic science, but true, notwithstanding that from the standpoint of the solar system the earth is rotating at approximately 1,000 miles per hour, orbiting the sun. When I walk at 2 miles an hour to pay for my coffee, it would be bizarre to claim that I am travelling at a rate of 1,002 miles per hour. Back to consciousness: the reality of consciousness (say, my awareness of an extreme toothache) does not depend on spatial frameworks or even conceptual (or philosophical) frameworks. I would have no less certainty in the reality of my toothache if I was assured that such agonizing pain would not be recognized from the standpoint of Dennett's naturalism. So, belief in the appearance of consciousness might be likened to the appearance that the earth is stationary insofar as both are true, while the second depends on the framework of measurement (establishing the compatibility of recognizing

stationary parts of earth and the movement of the earth itself), but the appearance of consciousness is so resilient that any framework that omits it has left out a vital, evident reality.

Second, more significantly, there is something vexing about using the terms "nonconscious" and "unconscious" interchangeably. The term "unconscious" is typically used either to describe persons who have lost consciousness or who have beliefs and desires that are repressed or not directly accessible to conscious awareness. A familiar Freudian claim would be that a patient has an Oedipus complex, an erotic desire for one's mother. In answer to Dennett's question of "why care," I do not care about the feelings of non-conscious objects, but I very much care about the unconscious feelings of my own and those around me. I earlier alluded to my friend Michael. I believe that we both enjoyed a reciprocal, loving friendship, but it would break my heart to learn (based on Michael's extensive psychoanalysis) that unconsciously he regarded me as a vile imposter whom he loathed. If the unconscious exists (and some philosophers are skeptical on the issue), I think we have abundant reasons to care about unconscious beliefs and desires.

Third, and most importantly, if we take up the question of why anyone should care whether they are conscious, I am somewhat baffled. I would rather be anesthetized during a major medical procedure which would otherwise cause me extreme pain. But in life itself, when we enjoy or endure friendships, romance, worthwhile work, show compassion and seek justice, play, make art, drive, vote, compete in sports, write and read, conscious awareness is indispensable.

What I believe Dennett's thought experiment brings to light is that if consciousness is an illusion, whether a machine is conscious is immaterial. But if it is not illusory, then it matters a great deal, both in our own case, in the case of relating to nonhuman animals, extraterrestrials (if there are any), God (if there is one) or other sacred realities (Brahman, Buddhahood, and more), and in relating to the machines we make.

Tentative Conclusion

If I am right about the evident reality and value of our being conscious selves (and thus right about the existence of consciousness), then we have reason to be suspicious of philosophies that eliminate or denigrate consciousness. While we therefore have reason to reject behaviorism and functionalism with respect to humans (and perhaps some nonhuman life), we may grant that behaviorism and functionalism would recognize that AI is intelligent. I would only add that this is profoundly different from recognizing AI as consciously self-aware beings.

The whole point of this paper is to bring to light the evident reality and importance of being conscious. As a Christian theist who has defended at length the reality of God as a supremely conscious living creator and sustainer of the cosmos, it is not my aim to argue that only biological embodied beings can be conscious. Indeed, it is my hope (along with billions of other religious practitioners) that this supremely conscious God will deliver us from what Seth describes as oblivion or nothingness.

I thank Kirk Allison, Jock McGreger, Greg Jessen, and members of the L'Abri Rochester community for fruitful dialogues about the nature of consciousness and AI.

Conflict of Interest

The author declares no competing interests.

References

Dennett, D. (2001). The zombie hunch: Extinction of an intuition? In A. O'Hear (Ed.), *Philosophy at the new millennium*. Cambridge University Press.

Dennett, D. (2013). *Intuition pumps and other tools for thinking*. W. W. Norton.

Habermas, G. (2023). Evidential near-death experiences. In A. Menuge, B. Rickabaugh, & C. Hasker (Eds.), *Minding the brain* (pp. 323–356). Discovery Institute Press.

Loose, J., Menuge, A. J., & Moreland, J. P. (Eds.). (2018). *The Blackwell companion to substance dualism*. Wiley-Blackwell. https://onlinelibrary.wiley.com/doi/book/10.1002/9781119468004

Lund, D. (2005). *The conscious self*. Humanity Books.

Sacks, O. (1985). *The man who mistook his wife for a hat*. Summit Books.

Seth, A. (2022). *Being you: A new science of consciousness*. Faber.

Taliaferro, C. (1994). *Consciousness and the mind of God*. Cambridge University Press.

Taliaferro, C. (2024). Mind and consciousness. In *St. Andrews Encyclopedia of Theology*. https://www.saet.ac.uk/Christianity/MindandConsciousness

Taliaferro, C., & Reasoner, P. (2026). *The meaning of life from sacred and secular perspectives*. Bloomsbury.

Yetter-Chappell, H. (2025). *The view from everywhere: Realist idealism without God*. Oxford University Press.

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

AI, Persons, Free Will, and a Policy Recommendation

John Lemos

McCabe Professor of Philosophy, Department of Philosophy, Coe College, USA.

jlemos@coe.edu

Abstract

This essay examines whether artificial intelligence (AI) could ever qualify as persons entitled to the same rights and responsibilities as human persons. The discussion engages two classic contributions to the philosophy of mind: William Lycan's argument that a suitably programmed machine could possess mentality and free will, and John Searle's Chinese Room argument, which casts doubt on the claim that syntactic symbol manipulation alone can generate genuine understanding or intentionality. The author concurs with Lycan that a properly programmed AI could, in principle, possess free will, but advances a distinct account of this possibility grounded in an event-causal libertarian theory of free will. The author argues that he and Robert Kane have both advanced theories of libertarian free will that could allow a properly programmed AI to possess free will. Nevertheless, the author contends that Searle's critique furnishes substantial grounds for skepticism regarding the personhood of any future AI, since genuine understanding and first-person intentionality may not be realizable in computational systems. Given this residual uncertainty, and given the profound moral and legal implications that would follow from granting personhood to AI, the essay proposes a precautionary policy recommendation: we should refrain from creating AI systems that, for all we know, may or may not be persons. The argument thus integrates metaphysical analysis of free will and personhood with applied ethical and policy considerations, offering a framework for responsible development of increasingly sophisticated artificial agents.

Original Research

Keywords:

artificial intelligence, personhood, free will, libertarianism, moral status, AI policy.

Received: 2026/03/12; **Received in revised form:** 2026/05/15; **Accepted:** 2026/06/08; **Published online:** 2026/06/12

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

Copyright © The Author

Introduction

We typically think that most adult human beings have capacities for conscious understanding – conscious knowledge – of themselves and their surroundings. We also think the typical adult human being has powers of reason and that they can act rationally in ways that make them responsible for what they do. And we typically think that to be responsible for what we do – or, at least, to be responsible in such a way that we could deserve praise or blame for our actions – we must also possess free will.

I believe it is possible that one day there may be AI that possesses such powers of conscious understanding, reason, and free will. I also believe that if we knew such AI existed, then we should regard them as persons and grant them the same rights we grant to human persons. In exploring the prospects for there one day being conscious, rational AI that possesses free will, I consider two old, classic essays by William Lycan and John Searle (Lycan, 1987/2013, pp. 358–363; Searle, 1980/2013, pp. 351–358). In the essays I discuss, they both express an openness to the possibility of there one day being AI that are persons, though Searle, as we will see, is – and I think rightly – more skeptical about this.

I start by discussing key points in Lycan's essay. I argue that he is correct in thinking that were there ever robotic AI who acted for all the world in the way that human beings do, then we should regard them and treat them as persons. He maintains that AI could even possess free will, and in doing so he explains how they could meet soft-determinist standards for free-willed action. While I agree that AI could one day possess free will, I give reasons to reject his soft-determinist account, and I go on to argue that there is no reason in principle why AI could not possess libertarian free will. In later parts of this article, I discuss Searle's essay, noting the grounds of his more skeptical view of the prospects for AI personhood.

Unlike Lycan, Searle is more concerned that the power for conscious understanding may be tied to having the kinds of human brains that we have. Searle thinks it may be that no man-made AI that lacks a human brain may be capable of any conscious understanding.

I argue that his deeper skepticism about the prospects of AI personhood is warranted, but I nonetheless maintain that if we get to a point where AI are indistinguishable from human beings in their behavior patterns, then we would be morally obligated to nonetheless treat them as persons with all of the rights and responsibilities of human persons. I conclude by arguing that such considerations morally justify a ban on making AI beings that blur the lines between persons and nonpersons.

William Lycan's More Optimistic Vision of the Prospects for AI Personhood

In his essay, William Lycan describes a fictional robot, Harry, who looks and acts like many human beings. In the following passage, he describes Harry and raises some interesting questions about him:

> Harry (let us say) is humanoid in form – he is a miracle of miniaturization and has lifelike plastic skin – and he can converse intelligently on all sorts of subjects, play golf and the viola, write passable poetry, control his occasional nervousness pretty well, make love, prove mathematical theorems (of course), show envy when outdone, throw gin bottles at annoying children, etc., etc. We may suppose he fools people into thinking he is human. Now the question is, is Harry really a *person*? Does he have thoughts, feelings, and so on? Is he actually conscious, or is he just a mindless walking hardware store whose movements are astoundingly like those of a person? (Lycan, 2013, p. 359)

Clearly, Harry is not a human being; he is not a member of the biological species *Homo sapiens*. But might he not, nonetheless, be a person, that is, a being with the kind of conscious understanding of himself and the world and powers of will and reason that make him deserve full moral consideration and, thus, make him deserving of rights to life and liberty which we ought to respect?

It is Lycan's view that were we ever to encounter a being like Harry, then we should regard him as a person and treat him with the same moral respect that we should show toward human beings. While Harry is not a human being and not made of the organic, biological materials we are made of, we should nonetheless assume he is a person and deserves our moral respect. In support of this view, Lycan talks about how we come to believe that other human beings besides ourselves are persons deserving of our moral respect. He notes that to each one of us it is self-evident that we have conscious understanding of ourselves and the world and powers of will and reason. For example, it is self-evident to me that these things are true of myself, and I presume it is self-evident to you, my dear readers, that these things are self-evidently known by you of your own selves. But, as Lycan claims, what is not self-evident to me, nor to any of you reading this, is that there are any other beings that have conscious understanding of themselves and the world and powers of will and reason. Rather, each one of us comes to know of the existence of other persons through rational inference.

It is Lycan's view that, to know that other human beings have conscious understanding of themselves and the world and the powers of will and reason

– to know that other human beings are persons – we must infer this based on the behavior of other human beings. The idea here is that when I have certain beliefs and desires, I act in certain ways, and when I see other human beings acting in the ways that I act when I have such beliefs and desires, I have reason to believe that they have similar beliefs and desires. For instance, when I am hungry and there is no food in the house, then I go out to secure some food to eat, typically by going to the grocery store. Because I act this way in such circumstances, then when I see that my wife hasn't eaten in the past twelve hours, and I see her look in the empty refrigerator and cupboards, and I see her walking down the street toward the market, I presume that she feels hungry, she knows there's no food in the house, and she is going to get some food. I presume she has similar kinds of conscious feelings, knowledge, and powers of reason that lead her to will her movement toward the market for food.

According to Lycan, since this is how we establish that there are human beings other than our own selves that have conscious understanding and powers of will and reason, and since Harry the robot also behaves in the ways that we ourselves would act given our beliefs, desires, feelings, and reason in similar circumstances, we then have good reason to believe that he, too, has powers of conscious understanding, will, and reason. That is, we have good reason to believe that he, like ourselves and other human beings, is a person. Lycan states:

> We know that a human being has such-and-such mental states when it behaves, to speak very generally, in the ways we take to be appropriate to organisms that are in those states. . . . We know for practical purposes that a creature has a mind when it fulfills all the right criteria. And by hypothesis, Harry fulfills all our behavioral criteria with a vengeance. . . . It follows that we are at least *prima facie* justified in believing him to be conscious.

We have not *proved* that he is conscious, of course – any more than you have proved that I am conscious. An organism's merely behaving in a certain way is no logical guarantee of sentience; from my point of view it is at least imaginable, a bare logical possibility, that my wife, my daughter, and my chairman are not conscious, even though I have excellent, overwhelming behavioral reason to think they are. But, for that matter, our "standard behavioral tests" for mental states yield practical or moral certainty only so long as the situation is not palpably extraordinary or bizarre. (Lycan, 2013, p. 360).

He goes on to note that many will think, and rightfully so, that the situation with Harry is "palpably extraordinary or bizarre," as he acts just like a human being, but he is not a human being. Thus, some will balk at the suggestion that he is a person with powers of conscious understanding, reason, and will.

Lycan notes that what makes Harry's case different is the fact that he is not born of human parents (rather, he was made in a lab) and that he is not made of the same things as human beings (rather, he is not made of carbon compounds but of silicon). But Lycan goes on to argue that these facts don't give sufficient reason to think that he lacks personhood, that he lacks the powers of conscious understanding, reason, and will.

I see no obvious way in which either a creature's origin or its sub-neuroanatomical chemical composition should matter to its psychological processes or any aspect of its mentality. (Lycan, 2013, p. 360)

With this in mind, he famously tries supporting this intuition with the example of Henrietta, who is born as a human being and who, as an adult, has her body parts – arms, legs, heart, even parts of her brain – gradually replaced with inorganic parts over a long period of time. He has us suppose that, eventually, over a long stretch of time, all her body parts, including all her brain parts, are replaced with inorganic (nonbiological) parts, but in the process "her intelligence, personality, perceptual acuity, poetic abilities, etc., remain just as they were before" (Lycan, 2013, p. 360). Were this to happen, it seems we ought to regard Henrietta as enduring through all this change and having all the same rights and responsibilities she had when she was composed of biological parts.

Interestingly, before concluding his essay, Lycan goes on to contemplate a possible objection to his view that nonbiological machines could conceivably one day rightly be regarded as persons. In doing so, he notes that some might say that if such machines are rightly considered to be persons, then they should have free will. He says it could be argued that, since any such machines would operate on programming given them by their makers, they would lack free will. The worry is that if all of the behavior of a machine, like Harry, is just the determined result of programming placed in him by his programmer, then he would lack the kind of free will that some might think is essential to personhood.

Now, Lycan could have simply rejected the notion that having free will is an essential ingredient of personhood. But he doesn't. Instead, he argues that there are ways in which machines, like Harry, could rightly be thought to have free will. In support of this, he argues that, for all we know, soft determinism about free will might be true, and if so, then the fact that Harry always acts and chooses in accordance with his programming does not establish that he lacks free will. Lycan says soft determinism is the view that "to have freedom of choice in acting is (roughly) for one's action to proceed out of one's own desires, deliberation, will, and intention, rather than being compelled or coerced by external forces regardless of one's desires or will" (Lycan, 2013, p. 361). Thus, on this conception of free will, even if all of Harry's desires, deliberations, willings, and intentions are the determined result of the way he is programmed

by his designers and his environment, he would still act with free will as long as his actions are caused by those desires, deliberations, willings, and intentions. After all, while his thoughts, desires, and deliberations may all be determined by the complex interplay of his programming and environmental influences, those thoughts, desires, and deliberations are still *his*, and if he acts in accordance with his desires, thoughts, and deliberations, then, on the soft determinist view, he is acting freely. Harry would only act unfreely when external forces inhibit his capacity to act in accord with his thoughts, desires, and deliberations. Since he embraces such a soft-determinist perspective on the nature of free will, Lycan sees the issue of free will as posing no threat to the idea that machines like Harry could be persons. Thus, he thinks that maybe one day in the future such nonbiological machines may come into existence, and if they do, we should treat them as persons, granting them rights to life and liberty.

There are two points I want to make about Lycan's essay. First, while I agree with Lycan that robots like Harry and Henrietta could possess free will, I don't think he has given an adequate account of how this is possible. Second, I agree that if robots, like Harry and Henrietta, came into existence, then we should treat them as persons, granting them all the same rights to life and liberty that we grant to human persons. However, unlike Lycan, I believe there are sufficient reasons to doubt the personhood of Harry and Henrietta, and I believe these reasons for doubt have moral implications for whether we should allow for the construction of such robots. In the next section of this essay, I will discuss issues of free will, arguing that soft-determinism is false and that even on a correct alternative theory of free will, the libertarian theory, it is conceivable that one day there may be robots that meet libertarian standards of free will. After that, I will consider ideas and arguments from John Searle which raise significant doubts about whether any robots will ever meet the conditions of personhood, and I will discuss some of the moral implications of these lingering doubts.

Free Will and Robots

In explaining how robots can have free will, Lycan appeals to a soft-determinist conception of free will. However, there are good reasons to reject such approaches to free will. On such a view of free will, when a person's actions are causally determined by their own beliefs and desires and deliberations, and are not the products of external force or coercion, then they act with free will. For instance, if I want to go into a swimming pool because I think it would be pleasant, and I believe it would not be wrong of me to go into the pool, and these beliefs and desires causally determine me to go into the pool, then I go into the pool of my own free will. But if I have no such beliefs

and desires, and if a strong wind blows me into the pool while I'd rather not go in, then I do not go into the pool of my own free will. If the soft-determinist theory of free will is true, then even if everything we ever do is a causally determined consequence of prior states and events, we can still have free will.[1] And, as noted, Lycan believes that since soft-determinism about free will is true, robots like Harry and Henrietta can have free will. That is, he believes that even if robots receive their programming from the engineers and programmers who design them, and even if all their behaviors are the determined result of the complex interplay between their programming and the situations they face in their environment, they could still rightly be said to possess free will.

I reject such views of free will. To have free will, we must have the kind of control over our lives and actions which grounds deserved praise for good acts and good moral character and deserved blame for bad acts and bad moral character. But meeting soft-determinist standards of free will does not suffice for this, and this is because we cannot have desert-grounding free will if everything that we ever do is causally determined. Rather, to have the kind of free will that is the ground of desert, some of our actions must be causally undetermined free-willed acts. Notice I say that *some* of our actions must be causally undetermined free-willed acts if we are to have free will. It is my view that many of our free-willed acts may be causally determined by our character or motivational states, but to be free-willed, such determined acts must issue from a character, a self, we have formed by prior causally undetermined free-willed acts. Only if this is the case can we possess desert-grounding responsibility for such causally determined actions. As did the great libertarian theorist of free will, Robert Kane, let us call these causally undetermined free-willed acts which shape our character, our self, "self-forming acts" (SFAs).

To see why free will requires the ability to form oneself through causally undetermined free-willed actions – to see why the capacity for SFAs is essential to free will – suppose that Harry has been programmed with certain

1. In contemporary philosophy, soft-determinists are typically understood as believing that (a) free will is compatible with determinism and (b) determinism is true. Thomas Hobbes ([1651] 1958), David Hume ([1748] 1974), and John Stuart Mill (1874) were soft-determinists. But in contemporary philosophy, not as many philosophers are committed to the truth of determinism, while many nonetheless still believe that even if determinism were true, or even if all human actions were determined, we could still have free will. Such philosophers tend to refer to themselves as "compatibilists." For some recent and contemporary defenses of compatibilism, see A. J. Ayer (1954), Daniel Dennett (1984a, 1984b, 2003), Owen Flanagan (2002, ch. 4), Harry Frankfurt (1969, 1971), G. E. Moore (1912), Kai Nielsen (1971), J. J. C. Smart (1963), Gary Watson (1975), and Susan Wolf (1987, 1990). John Martin Fischer defends what he calls "semi-compatibilism," the view that while free will is not compatible with determinism, moral responsibility is. See Fischer (1994, 1999, 2006) and Fischer and Ravizza (1998).

sorts of values. Perhaps he is programmed with a belief in the Kantian formula of the end, which says we ought always to treat persons as ends in themselves and never as mere means. Thus, in any given situation involving the treatment of persons, Harry always treats them as ends and never as mere means. Thus, he never lies, nor cheats, nor steals, nor rapes, nor murders, because he is programmed to act in accord with the principle of ends. He follows these Kantian norms even when he knows that he could get away with and profit from violating these norms, and, again, he follows these norms because this is how his programming operates. Suppose Harry always does the right thing as a result of his being given this Kantian programming, which determines that he will always make the ethical choice. Would he deserve our praise for always doing the right thing? Many people would, I think, say, "No, he does not deserve our praise, because all of his right actions are just the necessary consequence of the programming that has been bestowed upon him." And I agree with this sort of reaction. Furthermore, if Harry had been programmed with a bad moral framework which necessitated his acting badly in many situations, then I do not think he would deserve our blame either.

Now, if I am right that having free will grounds deserved praise and blame for our actions, then soft-determinism is misguided. For if soft-determinism is true, then Harry should be said to act with free will whenever he acts in accordance with his beliefs, desires, and deliberations. But this would entail that he acts with free will in acting with the Kantian morals he has been programmed to have and which he necessarily follows. Yet, it seems intuitively obvious that he doesn't deserve praise for all the good actions he does in this manner, nor would he deserve blame if he had been programmed to act badly. Thus, if it's true that free will grounds deserved praise and blame, then soft-determinism seems to be false.

To have desert-grounding free will, agents – and Harry included – must be responsible for shaping their own moral characters, their own selves. And to do this, they cannot act in determined ways from programs or characters imposed upon them solely by external factors, whether by programmers, enculturation, genetics, or the like. Rather, there must be moments in their lives in which they make causally undetermined free choices which have an impact on shaping their future selves for better or for worse. The sort of view that I am defending here is a libertarian view of free will of the sort that Robert Kane and I have defended over the years in many of our writings (Kane, 1996, 2005; Lemos, 2018, 2021, 2023a, 2023b). My own views are largely Kaneian in nature; however, we offer slightly different views about the nature of self-forming acts.[1] I would note that Kaneian

1. For expressions and defenses of my own modified view of self-forming acts (SFAs), see my (2018, Ch.5, esp., pp.104-112, 2021, 2023a, Ch. 6, 2023b, 2025a, 2025b). For critical

views of free will raise certain sorts of questions that must be answered to make the theory plausible. In what follows, I will say more about this and then move on to talk about what must happen in a robot if it is ever to act with desert-grounding free will. In discussing these issues further, I will just focus on Kane's view, not my own variant of Kane's view.

Kane maintains that, to be responsible for our characters – to deserve praise or blame for our characters and for the actions causally determined by our character – we must be responsible for causally undetermined actions which shape our character. For instance, if I have a bad character which frequently leads me in a causally determined way to do bad things, then, on Kane's view, to be responsible for those causally determined bad acts and to deserve blame for those bad acts, I must have contributed in significant ways to forming that bad character through prior causally undetermined free-willed acts. Many philosophers are skeptical about whether a person can have desert-grounding control over causally undetermined actions. In support of this skepticism, it has been argued that if an action is causally undetermined at the moment of choosing, then two persons, P and P*, in different possible worlds facing the same choice between two options, A and B, could choose differently even if everything about them, including all their actions and psychological states, had been exactly the same until the moment right before their choice. But if their life histories and their psychological states had been exactly the same throughout their entire lives right up to the moment before they choose, and they choose differently, then nothing about them explains the difference in their choices. As such, their different choices seem to be unexplainable random happenings, mere matters of luck, suggesting that they would lack desert-grounding control over such causally undetermined choices.[1]

Kane famously maintains that we can be responsible for undetermined actions, including undetermined choices. In support of this point, he appeals to Austin-style examples, so named after the famous philosopher J. L. Austin, who originally presented such examples. Kane has us consider an assassin who is trying to shoot and kill a prime minister. He has us imagine that, as the assassin takes aim and is pulling the trigger, it is causally undetermined whether the assassin will have a nervous twitch making his shot go off course, leading

discussions of my modified view of SFAs, see Caruso (2025), Haji (2022, 2023) and Kane (2024, Ch.7, pp.175-181), Pereboom (2025), and Sekatskaya (2025).

1. This possible worlds version of the luck objection has, perhaps, been most forcefully defended by Alfred Mele (1998, 1999a, 1999b). Another famous version of the luck objection is Peter van Inwagen's "replay" or "rollback" argument (2002). For other expressions of luck objections, also see Allen (2005), Caruso (2021), Clarke (2002), Haji (1999, 2005), Levy (2011), Pereboom (2001, 2007a, 2007b, 2014), Strawson (2000), Waller (1988).

him to miss the prime minister. Kane says suppose there is no nervous twitch and the assassin succeeds in killing the prime minister. In such a case, the assassin's killing of the prime minister would be causally undetermined, but he would still be responsible for killing him. And, according to Kane, he would be responsible for killing the prime minister because he would have succeeded in what he was trying to do. When we succeed at what we are trying to do, we will be responsible for what we do even if what we do is causally undetermined.

According to Kane, this helps explain how we can be responsible for causally undetermined free choices. To see this, let's first consider the kinds of conditions that give rise to the making of causally undetermined self-forming choices. Kane maintains that these kinds of choices occur when we are torn between two or more different possible courses of action, A and B, and we have some significant motivation to do either one. These are the kinds of situations that lead us to deliberate prior to choosing which course of action to take. In the most recent expression of his view,[1] Kane maintains that during deliberation, as the agent finds a certain option, say A, more attractive, the agent will likely start making an effort of will – start trying – to choose A. However, since he is strongly attracted to option B as well, it is causally undetermined whether the agent's effort to choose A will succeed. If that first effort to choose A succeeds, leading him to choose A, then his choice will be causally undetermined, but he will be responsible for it, because his subsequent choice of A will be a result of his effort. However, if his initial effort to choose A fails, then he may either stop deliberating and not choose, or he may deliberate further, possibly making another effort to choose A or making an effort to choose B. Suppose he makes an effort to choose B. It will then be undetermined whether this effort succeeds, because he still has some significant attraction to choose A. But if this later effort to choose B succeeds, then, again, he will be responsible for his choice of B, as it is something he tried to choose.

On Kane's view, when making these undetermined choices, whatever the agent chooses will be a result of his effort to make that particular choice. With our example in mind, if he chooses A, that will be the result of an effort to choose A; if he chooses B, that will be a result of his effort to choose B. And since, as the Austin-style examples demonstrate, we are responsible for

1. In very recent years, Kane has changed his understanding of SFAs. He used to believe that in SFAs the agent simultaneously makes distinct efforts of will to choose each option considered in deliberation. Due to criticisms of this point, he now rejects the idea that agents engage in such simultaneous conflicting efforts of will. For expressions of the relevant criticisms of Kane's old view, see Balaguer (2010, 179n23), Caruso (2021, 49), Clarke (2003, 88), Ekstrom (2003), and Lemos (2018, 73-81, 2023a, 119-127). For expressions of Kane's more recent view, which is designed to avoid these problems as well as the problem of luck, see Kane (2021, 2024).

undetermined acts when we succeed at what we are trying to do, we are then responsible for causally undetermined free choices in cases like those described above.

Furthermore, it is Kane's view that in our lives we often face these choices when we are torn between doing what is morally right and doing what is wrong, or when we are torn between satisfying some present desire when we would be better off waiting and pursuing some other course of action. And he believes that, depending on how we choose in these sorts of cases, it contributes to the formation of our character, our selves, contributing in significant ways to whether we are morally decent persons or morally bad persons or, for that matter, weak-willed or strong-willed. The idea is that if I commit too many self-forming acts in which I make bad choices, then I am likely contributing to my having a morally bad character. Thus, when in later years I have a bad character and this determines me to act badly, then I am responsible for those later bad actions because I have contributed in significant ways to making myself morally bad.

I am very sympathetic with this picture of the nature of free will and responsibility. While I have developed a slightly different view about the nature of self-forming acts, replacing Kane's efforts of will directed at particular choice options with the indeterministic assigning of evaluative weight to different choice options, I find the overall framework of his theory and the grounds for his rejection of soft-determinist and compatibilist views of free will to be correct. With these points in mind, let's consider what must be the case if a robot, an AI, is to have free will in the Kaneian sense. First, the robot must be constructed and programmed in such a manner that it has conscious understanding of itself and the world. It must be able to formulate goals and be able to think and reason and be consciously aware of its goals and its reasoning. Second, it must be able to make efforts of will to achieve its goals, since, on Kane's view, making such efforts of will is crucial to making causally undetermined free-willed choices. Third, there must be situations it faces in which the goals it has come into conflict, creating the need for deliberation and choice among these conflicting goals. And fourth, there must, of course, be causal indeterminacy at work in some of the practical reasoning – some of the deliberations and decision-making – of any robot that has free will.

Regarding the last point about causal indeterminacy in deliberation and decision-making, it is important to understand that, with respect to human beings, when an undetermined free-willed choice between two options, A and B, is made, it is not sufficient for the agent to make efforts for the different choice options and then merely have one effort succeed merely due to a randomly occurring neural blip. Rather, Kane has always been clear that the causal indeterminacy of neural

functioning that allows for undetermined free choices opens up a window in which an agent can make one of his efforts of will overcome resistance, allowing the agent to make the choice to which that effort was directed. So, for instance, if I choose A and not B, the indeterminacy in neural processing merely leaves it open as to whether my effort to choose A will succeed. The choice of A must not be seen merely as the result of some random neural blip that causes that choice. Rather, the indeterminacy in neural functioning merely allows that my effort to choose might or might not succeed, and when the effort does succeed, the choice is a result of that effort, and I am responsible for the choice. As such, the indeterminacy merely opens up a window in which I can make, in an undetermined way, one of my efforts succeed at making a choice, thereby establishing my responsibility for the choice.[1] I stress this point here to make clear that in any AI that has free will, it would not suffice for its undetermined choices *merely* to be the product of random decision -output generators. Were this the case, the choice would not be free-willed. Rather, for an AI to make a free-willed undetermined choice of A instead of B, the AI must make an effort of will to choose A, and it must be able to make that effort succeed in the face of resistance brought on by its recognition that option B is also a valued goal that it has.

---

1. As some evidence of this aspect of Kane's view, I would note how Kane has replied to the idea that on his view our undetermined free-willed decisions would exert no more free will than having our decisions decided by random decision output generators. Dennett (2003, pp. 132-133) has made such an objection, but he notes that in correspondence Kane replies:

   > The indeterminacy-producing mechanism [leading a person to make an undetermined free-willed choice] must be responsive to the dynamics within the agent's own will and not override them or it would be making the decisions and not the agent. [Quoted from Kane in Dennett (2003, 133)]

   Elsewhere when discussing SFAs, Kane writes:

   > [O]ne needs to say more than that certain reasons (for doing A) win out in the process of deliberation at the moment of choice, or certain other reasons (for doing B) win out, and it is undetermined which ones win out. One must also be able to say that the agent makes one set of reasons win out over the other at the moment of choice, so that the agent can be fully responsible for causing it to be the case that one choice rather than the other is made, despite the indeterminism. And this requires more resources.
   >
   > To do it, I argue, the activity of the agent prior to choice must not merely be imagined to be a single process (deliberation) with multiple possible outcomes (deciding to A or deciding to B), one of which occurs by chance. One must also consider that there is an internal complexity to the preceding process of deliberating that gives the agent more of an active role in settling which reasons win out over the others. This can be done if, instead of merely imagining a single process (deliberation) with multiple possible goals, one also imagines that within that larger process, there are multiple processes, each with a specific goal the agent is trying to realize. These processes, of course, are my efforts of will or volitional streams…
   >
   > In such a scenario, it would not merely be the case that one set of reasons "wins out" over the other, but rather that the agent makes one set of reasons prevail over the other by making an effort to do so against the competing effort to make a contrary choice. (Kane, 2014, p. 208).

Kane's theory of free will is a naturalistic theory of free will. He makes no appeal to Kantian noumenal selves or Cartesian nonphysical souls to make sense of free will. Rather, on his view and mine, human beings are complex material beings. Our thinking, understanding, reasoning, and willing arise from the operations of our brains. Thus, if the complex material beings that we are can engage in causally undetermined free-willed decision-making, then I see no reason *in principle* why there may not someday be machines, AI, that can do the same. This, of course, is not to say that there are any currently existing AI that have the kind of free will that makes them responsible for what they do. I think we are far from that.

In the next section, I will address some epistemological concerns about whether we could ever really know that we are in the presence of AI persons deserving of rights to life and liberty. Before doing so, I want to address one other nuance about AI free will, responsibility, and deserved praise and blame. Suppose that a robot, like Harry, has just been made. And suppose in making him he was programmed to have: (i) certain goals and certain moral beliefs; (ii) a significant amount of understanding about his own self and the world around him; (iii) the ways and means to increase his understanding about himself and the world in which he lives; and (iv) the ways and means to rationally deliberate in the face of any situations creating conflicts in his goals. What I want to note here is that Kane and I would say that when first brought into existence, Harry would have very little responsibility for anything that he does. Recall that, on the view I have put forward, we only engage in causally undetermined free-willed choices when we encounter conflicts in our will, and much of what we do is not done in the face of conflict. Rather, much of what we do is the determined result of the kind of character we have and our motivational states. Thus, when first brought to life and released into the world, Harry will mostly engage in causally determined acts programmed into him by his programmers. He would not be responsible for such causally determined acts. Rather, his programmers would be responsible for those acts. But in time, as he navigates his way through the world, if he faces conflicts in achieving his goals and must reason through and make undetermined choices through efforts of will, he may then slowly build up increasing levels of responsibility for those undetermined choices and for the character shaped by these choices. Over the years, as he makes more and more causally undetermined free-willed choices, he will increasingly contribute to his own self-formation, and this will make him increasingly responsible for his character and the determined choices he makes as a result of his character in the later stages of his existence.[1]

1. For an alternative recent defense of the possibility of AI free will, see Christian List (2025). Like Kane and myself, List thinks to have the power of free will agents must be able to make

John Searle's More Skeptical Thoughts on the Prospects for AI Personhood

In another famous old essay, John Searle presents an argument which raises significant doubt about whether there could ever be AI that have conscious understanding of themselves, their environment, and what they are doing. This is his classic Chinese Room argument. This argument proceeds by analogy. Searle has us imagine that he does not understand the Chinese language at all. He is put into a room with no windows, and there is one wall of the room where there is a slot through which written texts may be submitted. There is another slot where written texts may be sent out of the room. He has us also imagine that Chinese people are outside the room and that they are encouraged to submit Chinese written comments and questions into the room through the input slot. The Chinese people do this. Searle, inside the room, is supposed to provide responses to these written Chinese texts. In the room is a massive book with instructions telling the reader what Chinese texts to send out in response to any Chinese texts coming in. Searle uses these instructions to make replies to any Chinese texts which are sent into the room, and he sends out written Chinese responses through the output slot. The Chinese people outside receive the replies that Searle sends out. The replies that Searle sends out are perfectly sensible replies to any comments or questions submitted. For instance, if a message sent in Chinese into the Chinese room means "Hello, how are you?", Searle, using his response instruction book, sends out Chinese words which mean something like, "Fine, thank you. How are you?" But when Searle does this, he has no idea what his written responses say. In the room, Searle sends out all kinds of reasonable Chinese responses to Chinese comments and queries, but Searle, not understanding any Chinese, has no idea what the comments and queries coming in mean, and he has no idea what the comments and queries going out mean. The people outside the room will likely, and somewhat reasonably, conclude that there is someone inside the room who understands Chinese and understands their submitted comments and queries and understands the replies he sends out of the room. But they are mistaken, because Searle understands no Chinese at all.

Searle says that this is how computers operate. They receive inputs of data and give responses, but in doing so they have no understanding of the incoming

---

decisions while having alternative possibilities. But the way in which he makes sense of such alternative possibilities has been subjected, and I think rightly, to significant criticism. See Kane (2024, pp. 232-235) and Pereboom (2023).

For a more negative assessment of the prospects for AI free will, see Lynne Rudder Baker (1981), who argues that AI cannot act at all. A being that cannot act cannot have free will. For a critical assessment of Baker's view, see Berkich (2018). For a different account of the nature of agency and the prospects of AI agency, see Helen Steward (2012, esp. p. 91).

data, and they have no understanding of their responses to this incoming data. It is his view that computers are programmed to respond to data they encounter in the ways that their programmers want them to respond, but in doing this, the computers have no understanding of the incoming data nor of their responses to the data. Indeed, a highly skilled programmer might be able to program a computer to perfectly simulate an intelligent human being who can engage in long, extended conversations with a human interlocutor in such a way that, if the human being didn't know any better, he would think that he was talking with another human being. However, if a computer is programmed to do this, it would still not understand any incoming data or the meaning of any of its responses to this incoming data. For these reasons, Searle believes that no computer understands what it is or what it does, even though it may perfectly simulate human responses to its environment, including responses to questions and comments from human interlocutors (Searle, 2013, pp. 352–353).

To be clear, Searle's point is that computers process information – take in data and respond to data – by instantiating a program. And, given the right sort of program, they can *simulate* human consciousness and understanding, just as the set-up of the Chinese Room can simulate human consciousness and understanding of the Chinese language. But *merely* instantiating a program that simulates human consciousness and understanding does *not* suffice for actually being conscious and understanding (Searle, 2013, p. 355). According to Searle, computers, understood in this sense, can *never* possess consciousness and understanding. As such, computers, understood in this sense, could never be persons and could never possess free will nor be morally responsible for what they do, and we could have no obligations to any computers understood in this sense.

However, Searle is open to the possibility that an artifact – a man-made machine – can think. In answer to the question of whether a man-made machine can think, he writes:

> Assuming it is possible to produce artificially a machine with a nervous system, neurons with axons and dendrites, and all the rest of it, sufficiently like ours, again, the answer to the question seems to be, obviously, yes. If you can exactly duplicate the causes, you could duplicate the effects. And indeed it might be possible to produce consciousness, intentionality, and all the rest of it using some other sorts of chemical principles than those human beings use. It is, as I said, an empirical question. (Searle, 2013, p. 355)

But he continues, writing:

> "Ok, but could a digital computer think?"

> If by "digital computer" we mean anything at all that has a level of description where it can correctly be described as the instantiation of a computer program, then again the answer is, of course, yes, since we are the instantiation of any number of computer programs, and we can think.
>
> "But could something think, understand, and so on *solely* by virtue of being a computer with the right sort of program? Could instantiating a program, the right program of course, by itself be a sufficient condition of understanding?"
>
> This I think is the right question to ask, though it is usually confused with one or more of the earlier questions, and the answer to it is no.
>
> "Why not?"
>
> Because formal symbol manipulations by themselves don't have any intentionality; they are quite meaningless; they aren't even symbol manipulations, since the symbols don't symbolize anything. (Searle, 2013, p. 355)

Interestingly, Searle says, "Yes, a digital computer can think." But in reply to the question, "Could instantiating a program, the right program of course, *by itself* be a sufficient condition of understanding?" [my emphasis added], he says "No." His point is that what a thing is made of, or at least how it is physically organized, is crucial to whether it can be conscious, whether it can have thought and understanding. Just because there are currently computers that can be programmed to fool us into thinking that they are persons with understanding is not sufficient evidence that they actually do think and understand. According to Searle, for something to think and understand, to have genuine consciousness, it must be made of the right sort of material and organized in the right sort of way.

The passages from Searle which I quote above suggest that he is open to there being something materially different from us which has consciousness. He says, "[I]t might be possible to produce consciousness, intentionality, and all the rest of it using some other sorts of chemical principles than those human beings use. It is, as I said, an empirical question." But in the end, he suspects that the capacity for conscious awareness and understanding is a biological phenomenon. In a later passage in the same essay, he writes:

> Could a machine think? My own view is that only a machine could think, and indeed only very special kinds of machines, namely brains and machines that had the same causal powers as brains. . . . Whatever else intentionality is, it is a biological phenomenon, and it is as likely to be causally dependent on the specific biochemistry of

> its origins as lactation, photosynthesis, or any other biological phenomena. (Searle, 2013, p. 357)

In this classic essay, Searle is, in essence, maintaining that: (1) mere excellent simulation of human intelligence by a digital computer does not suffice for actually possessing intelligence; (2) a nonbiological, man-made digital computer made of the right sorts of materials and organized in the right sort of way could possibly be intelligent; and (3) he harbors serious doubts that such a being will ever be made, because he suspects there is something about the biological material of which we are made that is essential to producing intelligence. In response to Lycan's story about Harry the Robot, Searle would harbor serious doubts about Harry's possession of any intelligence – conscious understanding – of what it is and its surroundings. Searle would believe it is most likely that Harry merely simulates understanding and personhood without possessing understanding and personhood, and this is because Harry is not made of the right sort of material that would give us good reason to think he has conscious understanding and personhood. Lycan, by contrast, would be much more open to the idea that Harry has conscious understanding of himself and the world and free will and that Harry is a person. In talking about Harry, as well as Henrietta, Lycan writes:

> We do not discriminate against a person who has a wooden leg, or a mechanical kidney, or a nuclear heart regulator; no more should we deny any human or civil right to Harry or Henrietta on grounds of their origin or physical makeup, which they cannot help. (Lycan, 2013, p. 360)

I would note, however, that even Lycan would harbor some level of skepticism about Harry's personhood. As I noted earlier, he says regarding Harry the robot, "We have not *proved* that he is conscious, of course – any more than you have proved that I am conscious. An organism's merely behaving in a certain way is no logical guarantee of sentience; from my point of view it is at least imaginable, a bare logical possibility, that my wife, my daughter, and my chairman are not conscious, even though I have excellent, overwhelming behavioral reason to think they are" (Lycan, 2013, p. 360). But despite this reservation about Harry, Lycan does not see this as good reason to think of him as a nonperson nor to deny him the rights that we grant to human persons.

Concluding Remarks: On the Ethics of Building AI like Harry and Henrietta

In this essay, I've discussed two old and famous essays that are relevant to the

question of whether there may one day be man-made, nonbiological AI that have the kind of conscious understanding, reasoning abilities, and free will that would make them persons. Both the essay by Lycan and the one by Searle express openness to this possibility, while Searle seems to be more skeptical about it than Lycan.

Along the way, I've said a good bit about the nature of free will and what it would take for an AI to have free will. I argued that it's conceivable that an AI could have the kind of robust libertarian free will that's essential to ground just deserts and responsibility. Some people may think having such free will is an essential feature of personhood and full-fledged membership in the moral community.

In closing, I would note that I believe that if we ever do make robotic AI, like Harry and Henrietta, who act for all the world like human beings and who would easily be mistaken for human beings, we should regard them and treat them as persons and grant them all the rights to life, liberty, property, and so forth that we grant human persons. However, as both Searle and Lycan point out, just because a being meets all our behavioral tests for being a person, this is no guarantee it is a person. This fact, coupled with the fact that man-made AI, like Harry and Henrietta, would lack the material, biological brains of human beings, justifies significant doubt about whether they would have the kind of conscious understanding of themselves and the world that is essential to having desert-grounding free will and personhood.[1]

Again, if we end up making AI that are like Harry and Henrietta, then we should grant them all the same human and civil rights as human beings. And why? Not so much because we'd know that they deserve such rights, but because we should err on the side of safety. It would be horrible to deny them rights to life and liberty if they have mental lives that include the capacities for conscious understanding, reasoning, hoping, dreaming, fearing, and free will that are the hallmarks of personhood. At the same time, however, and as noted, there would be significant reason to doubt their personhood despite their certainly seeming to have it. Furthermore, regarding them as persons with all the rights and responsibilities that come along with this, we would, as a society, have to grant significant amounts of time, money, and effort in seeing to it that they are cared for and protected in the ways we care for and protect human

1. Just to be clear, regarding the connection between conscious understanding and free will, I would note that if a being lacks any conscious understanding, it will not have the kind of libertarian free will that Kane and I endorse. For instance, on Kane's view to act with libertarian free will one must engage in efforts of will aimed at making particular choice options, and doing this requires conscious understanding of the options and understanding the aims of one's efforts.

beings. As noted, it would be horrible to not recognize them as persons with rights and treat them as such, if they *are* persons with rights. But it would also be a great waste of time, money, and effort to treat them as persons and to respect robot rights if, in reality, they are unconscious automatons. Thus, with all these points in mind, I do not think that human beings should make AI that may lead us to wonder if they are persons with rights. We should not make AI, like Harry and Henrietta. We don't need them, as the world has plenty of persons as is. But if we do make AI, like Harry and Henrietta, who, for all we know, are persons, then, to err on the side of safety, we should treat them as persons, even though this may be a costly waste of our time, efforts, and resources.

I am, in essence, suggesting that once we have AI, like Harry and Henrietta, which (who?) for all the world – act like human beings in all the ways that would suggest they are persons deserving of moral rights and respect, we will then still have legitimate uncertainties as to whether they really merit the granting of such rights and respect. The problem faced here is well discussed in an important jointly authored report, *Taking AI Welfare Seriously* (Long et al., 2024), which was supported by Eleos AI Research and NYU's Center for Mind, Ethics, and Policy. In Section 1.2, "Risks of Mishandling AI Welfare," of this report, the authors do a wonderful job of spelling out this problem, noting that it would be terrible to not respect the rights of AI if they really are persons, but it would also be a terrible waste of resources, time, and effort to grant them rights if they ultimately are more like mindless automatons. As I argue in this essay, I think we should not make such highly sophisticated AI so as to avoid getting caught in between this rock and this hard place. Of course, as a society, we will have to address issues regarding at what point an AI is sufficiently person-like to ban its construction. I have not addressed this issue here, but it is one that must be confronted.[1]

▣ Conflict of Interest

The author declares no competing interests.

1. For a really nice discussion of issues dealing with this, see Lawrence & Morley (2021).

References

Allen, R. F. (2005). Free will and indeterminism: Robert Kane's libertarianism. *Journal of Philosophical Research, 30*, 341–355. https://doi.org/10.5840/jpr20053042

Ayer, A. J. (1954). Freedom and necessity. In *Philosophical essays* (pp. 3–20). St. Martin's Press. https://doi.org/10.1007/978-1-349-00132-3_12

Baker, L. R. (1981). Why computers can't act. *American Philosophical Quarterly, 18*, 157–163.

Balaguer, M. (2010). *Free will as an open scientific problem*. MIT Press.

Berkich, D. (2018). Machine intentions. *APA Newsletter on Philosophy and Computers, 18*, 3–10.

Caruso, G. (2021). *Rejecting retributivism: Free will, punishment, and criminal justice*. Cambridge University Press.

Caruso, G. (2025). Event-causal libertarianism, indeterministic weightings, and the pragmatic approach. *Philosophia, 53*, 905–914. https://doi.org/10.1007/s11406-025-00813-0

Clarke, R. (2002). Libertarian views: Critical survey of noncausal and event-causal accounts of free agency. In R. Kane (Ed.), *The Oxford handbook of free will* (pp. 356–385). Oxford University Press.

Clarke, R. (2003). *Libertarian accounts of free will*. Oxford University Press.

Dennett, D. C. (1984a). *Elbow room*. MIT Press.

Dennett, D. C. (1984b). I could not have done otherwise – so what? *Journal of Philosophy, 81*, 553–567.

Dennett, D. C. (2003). *Freedom evolves*. Viking Press.

Ekstrom, L. (2003). Free will, chance, and mystery. *Philosophical Studies, 113*, 153–180. https://doi.org/10.1023/a:1023940209581

Flanagan, O. (2002). *The problem of the soul: Two visions of the mind and how to reconcile them*. Basic Books.

Frankfurt, H. (1969). Alternate possibilities and moral responsibility. *Journal of Philosophy, 66*, 829–839. https://doi.org/10.2307/2023833

Frankfurt, H. (1971). Freedom of the will and the concept of a person. *Journal of Philosophy, 68*, 5–20. https://doi.org/10.2307/2024717

Haji, I. (1999). Indeterminism and Frankfurt style examples. *Philosophical Explorations, 2*, 42–58.

Haji, I. (2005). Libertarianism, luck, and action explanation. *Journal of Philosophical Research, 30*, 321–340. https://doi.org/10.5840/jpr20053041

Haji, I. (2022). Libertarianism and luck. *Journal of Philosophical Theological Research, 24*, 115–134. https://doi.org/10.22091/JPTR.2022.8297.2723

Haji, I. (2023). Further reflections on Lemos's indeterministic weightings view of libertarian free will. *Journal of Philosophical Theological Research, 25*, 119–130. https://doi.org/10.22091/JPTR.2023.9722.2929

Hobbes, T. (1958). *Leviathan*. Bobbs-Merrill. (Original work published 1651)

Hume, D. (1974). An enquiry concerning human understanding. In *Locke, Berkeley, and Hume, The empiricists*. Anchor Books. (Original work published 1748)

Kane, R. (1996). *The significance of free will*. Oxford University Press.

Kane, R. (2002). Some neglected pathways in the free will labyrinth. In R. Kane (Ed.), *The Oxford handbook of free will* (pp. 406–437). Oxford University Press.

Kane, R. (2007a). Libertarianism. In J. M. Fischer, R. Kane, D. Pereboom, & M. Vargas (Eds.), *Four views on free will* (pp. 5–43). Blackwell.

Kane, R. (2007b). Response to Fischer, Pereboom, and Vargas. In J. M. Fischer, R. Kane, D. Pereboom, & M. Vargas (Eds.), *Four views on free will* (pp. 166–183). Blackwell.

Kane, R. (2011). Rethinking free will: New perspectives on an ancient problem. In R. Kane (Ed.), *The Oxford handbook of free will* (pp. 381–404). Oxford University Press.

Kane, R. (2014). New arguments in debates on libertarian free will: Responses to contributors. In D. Palmer (Ed.), *Libertarian free will: Contemporary debates*. Oxford University Press.

Kane, R. (2021). Making sense of a free will that is incompatible with determinism: A fourth way forward. *Journal of Philosophical Theological Research, 23*, 5–28. https://doi.org/10.22091/JPTR.2021.7247.2595

Kane, R. (2024). *The complex tapestry of free will: A philosophical odyssey*. Oxford University Press.

Lawrence, D., & Morley, S. (2021). Regulating the Tyrell Corporation: The emergence of novel beings. *Cambridge Quarterly of Healthcare Ethics, 30*, 421–434. https://doi.org/10.1017/S0963180120000973

Lemos, J. (2018). *A pragmatic approach to libertarian free will*. Routledge.

Lemos, J. (2023a). *Free will's value: Criminal justice, pride, and love*. Routledge.

Lemos, J. (2023b). Haji and the indeterministic weightings model of libertarian free will. *Journal of Philosophical Theological Research, 25*, 101–118. https://doi.org/10.22091/JPTR.2023.9648.2920

Lemos, J. (2025a). Precis of *Free will's value: Criminal justice, pride, and love*. *Philosophia, 53*, 905–914. https://doi.org/10.1007/s11406-025-00870-5

Lemos, J. (2025b). Replies to Caruso, Elzein, Pereboom, Sekatskaya, and Smilansky. *Philosophia, 53*, 973–993. https://doi.org/10.1007/s11406-025-00904-y

Levy, N. (2011). *Hard luck: How luck undermines free will and moral responsibility*. Oxford University Press.

List, C. (2025). Can AI systems have free will? *Synthese, 206*, 115. https://doi.org/10.1007/s11229-025-05209-x

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* (arXiv:2411.00986) [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2411.00986

Lycan, W. (2013). Robots and minds. In J. Feinberg & R. Shafer-Landau (Eds.), *Reason and responsibility: Readings in some basic problems of philosophy* (pp. 358–363). Wadsworth. (Original work published 1987)

Mele, A. (1998). Review of Robert Kane's *The significance of free will*. *Journal of Philosophy, 95*, 581–584. https://doi.org/10.5840/jphil1998951122

Mele, A. (1999a). Kane, luck, and the significance of free will. *Philosophical Explorations, 2*, 96–104. https://doi.org/10.1080/13869799908520970

Mele, A. (1999b). Ultimate responsibility and dumb luck. *Social Philosophy and Policy, 16*, 274–293. https://doi.org/10.1017/S0265052500002478

Mill, J. S. (1874). *A system of logic*. Harper and Row.

Moore, G. E. (1912). Free will. In *Ethics*. Oxford University Press.

Nielsen, K. (1971). The compatibility of freedom and determinism. In *Reason and practice*. Harper and Row.

Pereboom, D. (2001). *Living without free will*. Cambridge University Press.

Pereboom, D. (2007a). Hard incompatibilism. In J. M. Fischer, R. Kane, D. Pereboom, & M. Vargas (Eds.), *Four views on free will* (pp. 85–125). Blackwell.

Pereboom, D. (2007b). Response to Kane, Fischer, and Vargas. In J. M. Fischer, R. Kane, D. Pereboom, & M. Vargas (Eds.), *Four views on free will* (pp. 191–203). Blackwell.

Pereboom, D. (2014). *Free will, agency, and meaning in life*. Oxford University Press.

Pereboom, D. (2023). Review of *Why free will is real*? *Criminal Law and Philosophy, 17*, 229–234. https://doi.org/10.1007/s11572-022-09634-8

Pereboom, D. (2025). Event-causal libertarianism without settling. *Philosophia, 53*, 961–972. https://doi.org/10.1007/s11406-025-00914-w

Searle, J. (2013). Minds, brains, and programs. In J. Feinberg & R. Shafer-Landau (Eds.), *Reason and responsibility: Readings in some basic problems of philosophy* (pp. 351–358). Wadsworth. (Original work published 1980)

Sekatskaya, M. (2025). Manipulation, indeterminacy, and the luck problem: On Lemos' *Free will's value*. *Philosophia, 53*, 953–960. https://doi.org/10.1007/s11406-025-00878-x

Smart, J. J. C. (1963). Free will, praise, and blame. *Mind, 70*, 291–306.

Steward, H. (2012). *A metaphysics for freedom*. Oxford University Press.

Strawson, G. (2000). The unhelpfulness of indeterminism. *Philosophy and*

Phenomenological Research, 60, 149–155. https://doi.org/10.2307/2653435

Van Inwagen, P. (2002). Free will remains a mystery. In R. Kane (Ed.), *The Oxford handbook of free will* (pp. 158–177). Oxford University Press.

Waller, B. (1988). Free will gone out of control. *Behaviorism, 16*, 149–162.

Watson, G. (1975). Free agency. *Journal of Philosophy, 72*, 205–220. https://doi.org/10.2307/2024703

Wolf, S. (1987). Sanity and the metaphysics of responsibility. In F. Schoeman (Ed.), *Responsibility, character, and emotions* (pp. 45–64). Cambridge University Press.

Wolf, S. (1990). *Freedom within reason*. Oxford University Press.

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

Agency Attributions and AI Agency

Hajj Muhammad Legenhausen

Professor of Philosophy, The Imam Khomeini Education and Research Institute, Qom, Iran.
heidi@gmail.com

Abstract

The contemporary debate surrounding AI agency is frequently stalled by an essentialist preoccupation: attempting to discern whether an artificial system possesses genuine autonomy or traditional belief-desire agency. This paper argues that such an approach is a category error that obscures the practical realities of AI deployment. By shifting the discourse from an ontological investigation of machine agency to an "attributionist" framework, we can better understand how agency is constructed through socio-technical relationships rather than residing solely in algorithmic code. Drawing upon the theological concept of *fā'il bi-l-tajallī* – agency expressed through mediated manifestation – this paper posits that AI systems act as conduits for human intent and institutional structure. This perspective provides a necessary alternative to current technical debates. Furthermore, by engaging with recent scholarship on responsibility gaps – specifically by synthesizing Sven Nyholm's seminal identification of the moral vacuum in autonomous systems with Herman Veluwenkamp's granular analytic distinctions – we move toward a more precise taxonomy of accountability. Veluwenkamp shows that the so-called "responsibility gap" must be disaggregated into Epistemic Responsibility Gaps (ERG), relating to system complexity and explainability, and Control Misalignments (CM), which stem from institutional delegation failures. Ultimately, this paper reframes the "Alignment Problem." It is not a purely technical challenge of matching an objective function to a static goal, but a socio-technical endeavor of normative integration. An aligned AI is not one that achieves perfect independence, but one whose operations remain within the legible, revisable, and accountable bounds of human institutional intent. Through this attributionist lens, we demonstrate that resolving alignment concerns requires less focus on the internal "will" of the machine and greater focus on the distributed responsibilities that arise with the instantiation of these systems.

Original Research

Keywords:

agency, AI agent, reflex, action, autonomy, responsibility gap, agent by self-disclosure

Received: 2026/02/16; **Received in revised form:** 2026/04/12; **Accepted:** 2026/05/20; **Published online:** 2026/05/25

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.2026.17198.3657

Copyright © The Author

Introduction: What Is Agency?

Agency is the ability to perform actions. Being an agent is having this ability. Agency is essential to personhood. Persons do things. The converse is not the case. There is agency in some animals we do not count as persons. This approach pushes the question of agency to inquiries into what counts as an action. In contemporary action theory, it is usually assumed that actions must be intentional. [1] Intentionality distinguishes actions from involuntary movements. In Islamic philosophy, however, intentional agency is only one kind of agency. God is considered to be an agent by self-disclosure, *fāʿil bi-l-tajallī*, and not an intentional agent. [2] Here, I will not assume that actions have to be intentional, even if it is granted that human action normally involves intentionality. Clearly, actions need not be intentional for at least some non-human agents.

It is widely assumed that human actions are bodily movements that are caused by the right combination of beliefs and desires. This approach to action and agency has a number of drawbacks (aside from making divine agency incomprehensible, given a classical view of theism according to which God has no desires). It also makes AI agency impossible, given that AI agents do not have bodily movements. As an alternative, I will focus on the issue of agency attributions. An action must be performed by an agent in such a manner that the action can be attributed to the agent as something the agent did. By considering the conditions under which the performance of an action is attributed to an agent, we may outline several kinds of agency attribution, types of agency that are attributed, and consider the variety of things to which agency is attributed, including God, angels, humans, some animals, and some AI systems. There is also social agency that can be applied to groups of animals who form swarms, herds, and packs to protect their members and hunt. Human corporate agents include tribes, states, commercial conglomerates, orchestras, political movements, and governmental agencies. With respect to responsibility, AI agents may be more similar to corporate than individual agents, since responsibility for what AI agents do devolves to corporations, designers, developers, and users, as well as regulatory agencies.

There are a number of philosophical issues that I will not take up here. There is a debate about whether corporate actions are reducible to individual actions. I will not assume they are, but I will not try to argue that they are not. Likewise, I will be non-committal about whether agency reduces to justified agency attributions or to socially accepted agency attributions, and I will take

1. For an influential critical discussion of intention, see Anscombe (1959).
2. I discuss this in Legenhausen (2021).

no position on the relationship between agency and freedom of the will. [1]

By utilizing agency attributions instead of relying directly on a theory of the nature of agency, we will rely on scalar attributions, since those making the attributions may be more or less inclined to ascribe a given kind of agency to a system. Instead of imagining agency as a property that a system either has or lacks, consider a weighted set of capacities and relations due to which various kinds of agency attributions will be made. Different "agency markers" (goal-directedness, adaptivity, initiative, normative sensitivity, etc.) can vary independently, yielding different levels or types of agency (Zhu et al., 2026).

The question of whether some AI systems have agency turns out to be misplaced. It would be more appropriate to ask what kinds of agency are attributed to an AI system, and by whom. All the parties that use AI are called stakeholders. If our task is to understand the senses in which various kinds of agency are attributed to AI systems by various groups of stakeholders, we should avoid a philosophical characterization of agency that makes it true by definition that AI is not capable of any sort of agency. At the same time, we should not claim that AI has agency just because of the widespread usage of the term "AI agent." There are several factors that can serve as reasons for agency attributions, and these may be weighted and applied in different ways by different groups of stakeholders.

Instead of beginning with a theory about the nature of agency, we might reject the question "What is agency?" and replace it with several others:

- What are the types of agency that can be attributed to AI systems?
- What kinds of evidence justify attributions of the various types of agency?
- What is the range of AI capabilities within which agency attributions are made?
- How do the attributions of agency to AI systems compare with attributions of agency to other entities?

Actions and Reflexes

If agency is the ability to perform actions, as opposed to the ability to move, the difference between an action and a motion will aid our understanding of agency. In animals, the difference appears in how the animal is related to the movements of its body. According to Michael Tomasello's recent study of the

1. For a review of these topics see part 6 of Ferrero (2022).

emergence of agency in the course of biological evolution (Tomasello, 2022), the movement of an animal is an action when it is under the animal's control. Tomasello considers the question of how we can distinguish when a movement occurs as the result of some mechanism and when it is a controlled action. His analysis draws on cognitive psychology, which is not an especially promising route if we want to understand AI agency and AI agents do not possess psychological states. Nevertheless, we can transpose some of his claims from psychology to the behavioral observations that can support agency attribution.

Tomasello describes agency as a certain kind of capacity to behave in order to achieve goals. Not any sort of goal-oriented activity will be due to agency. A standard question in action theory is how to distinguish a self-regulating mechanism, like a thermostat, from an agent. In both cases, we need to be able to identify goal-directed activity. A goal attribution will be justified to the extent that it helps explain a pattern of behavior across circumstances – including adjustments, failures, and changes in the environment. It should also support some counterfactual expectations: if relevant circumstances had been different, we would expect the system to have adjusted its behavior in identifiable ways. This still is not sufficient to distinguish an agent from an HVAC (heating, ventilating, and air-conditioning) system controlled by a thermostat. We do not consider the HVAC system an agent because we can explain how the system works without any teleology. When the temperature gets too low or too high, the heat or the air conditioning turns on. The result is keeping the temperature of a room within a certain range. That is the goal of those who use the system to stay comfortable. It is not the goal of the thermostat (Tomasello, 2022, pp. 14–16). We shift from mechanical explanations to teleological ones when the mechanical explanations become so complex that they prevent human understanding of how the system works. Then we explain the behavior of a system as agential when it is flexibly directed toward goals through the utilization of feedback that enables the system to take a course that avoids obstacles and increases the likelihood of success in achieving a goal. We attribute agency when we observe behavior that requires this sort of explanation. The explanation through goals and feedback does not mean that there is no causal sequence of events that brought about the observed behavior. However, the causal sequence at the micro-level fails to enable us to understand what the animal is doing when agency is involved. If an animal twitches as a reflex to some stimulation, we can explain the twitch as a nervous reaction. When the animal uses skills to catch prey, nervous reactions are also in play; but the reactions cannot explain what is being done by the animal when goals, skills, threats, and opportunities are left out of the explanations. It is the kind of explanation needed to achieve understanding of behavior that justifies agency attributions, independent of whatever the phenomenology of agency

might involve in the agents. We do not have to know what it is like to be a bat in order to know what sorts of things bats do, and how these things are different from mere bat reflexes.

One might attempt to distinguish actions from reflexes by their relation to the intellect. Actions result from practical reasoning. Reflexes do not. While Aristotle held that animals do not possess intellect, he did admit that practical wisdom is attributed to them:

> … even some of the wild beasts are said to be practically-wise – those that appear to have a capacity for forethought about their life. (Aristotle, trans. 2014, 1141a).

Aristotle is not attributing agency and practical wisdom to the animals, because he denies that animals have intellect. However, he admits that the forethought exhibited by animals has led some people to attribute practical wisdom to them.

The distinction between voluntary and involuntary movements was recognized by Aristotle and in other ancient sources. Many centuries later, Ibn Sina, in the second book of his *Canon of Medicine*, describes "animal" functions of the body that facilitate immediate responses to external stimuli, and distinguishes between movements directed by the conscious intellect and those triggered automatically by sensory-motor pathways (Ibn Sina, 1999). Six centuries later, Descartes, in his treatise *L'Homme*, described ocular fixation and other bodily motions as the result of "animal spirits" flowing from the pineal gland through the nerves to muscles, with sensory traction on nerve filaments opening pores in the ventricles and thereby releasing more spirits – an automatic chain that can reinforce itself. It would take a couple more centuries for the physiological description of the "reflex arc" to be developed, although Descartes' German and English translators have interpreted his observations as indicating something like the modern concepts of reflex and feedback (Descartes, 1972, p. 98, n. 146).

Marshall Hall introduced the term "reflex arc" in 1833 for the path from stimulus through the nerves to movements in glands or muscles to produce a response; and he demonstrated that certain responses in decapitated animals could occur without any involvement of the brain. In the 1860s, Ivan Sechenov published studies of the nervous system in which self-regulation and the inhibition of reflexes were described.

In the 20th century, Norbert Wiener made the mathematical analysis of feedback loops a centerpiece of his cybernetics, the study of communication and control in living organisms or machines; and he illustrated this in an

explanation of ocular reflexes (Wiener, 2019, pp. 184ff.). Self-regulation through feedback loops would later become a key factor in the development of machine learning and artificial intelligence.

In physiology, the study of reflexes has continued to progress by differentiating the types of reflexes and the number of synapses involved, as well as the role of feedback loops and positive and negative reinforcement. Some reflexes, like breathing, can be brought under control. Even the pulse may be lowered through meditation. Agency is not attributed to reflexes when relatively limited components of the nervous system are involved.

For AI systems, we can find outputs that are generated by the system in an automatic way, like reflexes. AI reflexes can be contrasted with outputs that result from more complex processing that is designed to maximize expected success in a project set for the system by its users.

AI Systems and Agency Attributions

There is a formidable number of technical terms used to describe AI systems that draw on the language of agency: there are AI agents, agentic systems, and various levels of AI autonomy. These terms are defined in the literature of the science of AI. This complicates our method of trying to understand AI agency through agency attributions. When experts and users use the language of agency to describe a system, they need not be attributing the same thing to it by calling it an agent. This poses a serious challenge to the approach advocated here of trying to understand AI agency by examining relevant agency attributions, because the mere use of words like "agency," "agent," and "agentic" is not sufficient to indicate that it is agency that is being attributed. There is no philosophically or biologically useful sense of agency that could be applied to a clock just because we say that it does something, namely, that it tells us the time. The difficulty can be overcome by recognizing that different groups of stakeholders may use the terminology of agency in different ways. We should not assume a core meaning of agency that can be discovered merely by examining attribution patterns.

There are a number of different groups that may attribute agency to some AI systems: academics in the field of AI, developers, users, courts, oversight agencies, and corporations that buy or sell AI services. If we are concerned with the systems themselves rather than how we interact with them as users, algorithmic complexity might provide a more useful account of grades of agency than the reactive attitudes we direct toward persons as responsible agents and withhold from very young children and those who suffer from

debilitating mental illnesses or who cannot be considered to have the competence needed for responsibility because of trauma or other reasons (Strawson, 2008; Bagnoli, 2021). The kinds of social legibility that prompt reactive attitudes, such as blame, resentment, praise, and gratitude, will be more pertinent to user attributions of agency to AI systems than to system developers. AI assistants, companions, and other kinds of systems can give users the impression of agency, as if the systems had a will of their own, because the responses they provide are unpredictable while remaining relevant. Marketing also encourages users to interact with the systems as if they were persons. The result is not only that users attribute human-like agency to the systems; the users' own agency is extended through devices to the systems that enable them to do things that they otherwise could not. The downside of this is that over-reliance on AI systems may have deleterious effects on the ability to make informed choices and may debilitate the user's own agency.

For developers and designers, the term "AI agent" is reserved for systems that have abilities that go beyond those of a simple AI assistant. While an assistant needs continual interaction with a user, an AI agent can complete tasks on behalf of users or other systems without further human inputs by planning its own workflow and using available tools. Systems are said to be autonomous to the extent that they can perform tasks without human intervention. Agency is ascribed to AI systems by considering the limitations of the information and tools that can be accessed in order to produce responses to inputs from the environment or from users, to perform specific tasks, or to carry out larger projects. In AI theory, systems are described that represent inputs and compress information. Agency is attributed to a system when the system's algorithms permit it to achieve efficiency in its decision-making given the limitations on the complexity required for the ways in which the system is designed to interact with its environment. From an algorithmic information theory (AIT) perspective, an agent's information base can be viewed as a compressed representation of its environment. To be successful, agents must be able to identify patterns in the environment needed for the generation of a model that is sufficiently compact to be usable. Agency in a system relies on its ability to compress the raw data to which it has access. To the extent that an agent can carry out a task while minimizing the data it uses to do so, it will be an effective agent. The more "complex" the environment, the more sophisticated the agent's internal algorithm must be to remain effective.

In their acclaimed textbook on artificial intelligence, Stuart Russell and Peter Norvig (Russell & Norvig, 2020) describe the simplest AI agents, called reflex agents or reactive agents, as those that respond to inputs without planning, coordinating, or setting goals. They define cognitive agents as those

designed to perceive their environment, reason about it, and make decisions based on objectives. A rational agent is a type of cognitive agent that acts to maximize its expected utility, where utility is measured by the objectives set by the system's designers and the chances that a task will be satisfactorily completed. An ideal rational agent would calculate how to achieve maximum utility for any contingency; however, this would require impossibly excessive computational resources. Finite AI agency, therefore, requires a trade-off between accuracy in its responses and manageable complexity. The agent needs to select which information to compress and which can be safely ignored so that it can achieve an acceptable level of utility, or successful responses, within its computational limits.

On this view, agents are classified by the complexity of their algorithms, the inputs they can act on, and the responses or action outputs they provide. A simple agent will be designed to have a low-complexity policy, such as "if p, do a." Agents become more complex when their algorithms are used to process complex inputs that elicit high-dimensional action outputs.

Another group that is involved in the attribution of agency to AI systems is that of courts, regulatory organizations, and oversight commissions. Here, the focus is on how to assign liability for harms that result from AI use and what policies should be adopted by governments for the use and development of AI. The Organisation for Economic Co-operation and Development (OECD) is an intergovernmental organization with 38 member countries that has published the first intergovernmental standards for AI. [1] To simplify the OECD definition, an AI system is a machine-based system that receives inputs from which it infers a certain kind of output. AI agents are described by the OECD (OECD, 2026) as systems that perceive and act upon their environment with a degree of autonomy, using tools as needed to achieve specific goals and adapt to changing inputs and contexts. We will devote more attention to autonomy in the next section. In AI terminology, AI agents must not be confused with agentic AI. Agentic AI involves the orchestration of several AI agents. According to the OECD:

> … agentic AI refers to systems composed of multiple co-ordinated AI agents that can break down tasks, collaborate, and pursue complex objectives autonomously over extended periods. These systems are designed to operate in more open-ended, less predictable physical or virtual environments and to function with minimal human supervision. (OECD, 2026, p. 20).

1. https://www.oecd.org/en/topics/ai-principles.html

Before any further consideration of how agency attributions are made by regulatory institutions, another concept closely related to agency needs to be introduced: autonomy.

Autonomy and Alignment

While Kant introduced the idea of the autonomy of the will as fundamental to moral obligation (Kant, 1785/2011, Ak. 4:431), he also recognized a sense of autonomy that could be used for corporate agents, such as a university or a state (*civitas*), to the extent they enjoy institutional self-determination (Holzhey & Mudroch, 2005, p. 59). AI systems may be considered autonomous in a sense analogous to institutional autonomy. Systems may be considered autonomous to the extent that they select from alternative courses without the selection being explicitly determined through outside intervention.

In discussions of AI, autonomy is measured by the degree to which a system can operate without ongoing human intervention. This is not a Kantian autonomy of the will. AI autonomy does not imply moral agency. There is a danger that since AI systems are designed to interact with humans like persons, moral agency will be ascribed to them. When things go wrong through the use and development of AI systems, the systems will be blamed instead of those who profit from them, design them, or use them.

Systems are said to be autonomous to the extent that they are able to complete tasks without reliance on human interventions and when their actions are not fully predictable. As the algorithmic complexity of the policy of an AI agent increases, the agent's behavior becomes less "predictable" via simple observation of the environment. As the algorithmic complexity of a system increases, there is often a conflation of "predictive autonomy" – the system's ability to operate without human intervention – with "moral autonomy." However, given our attributionist framework, this is a category error. Autonomy in AI is merely the expansion of the domain of attribution of types of agency.

Unexpected behavior from autonomous AI systems that clashes with human values has given rise to what is called the "Alignment Problem." If an AI system lies or cheats, this is considered an alignment problem. There are various strategies about how to fix it. Rules can be imposed on the system from outside (top-down approaches), or steps can be taken so that the systems themselves may develop something like norms approved by developers, oversight panels, or courts (bottom-up approaches). As AI systems become more agentic (in the sense of agentic AI frameworks), the "Alignment Problem" transforms from a technical challenge of objective-function

matching into a broader, socio-technical challenge of normative integration. Alignment, therefore, is not the pursuit of a static goal but the ongoing process of calibrating a system's "normative sensitivity" to match the shifting expectations of its diverse stakeholders. In this view, an aligned AI is not one that lacks autonomy, but one whose autonomous actions remain within the legible, revisable, and accountable bounds of human institutional intent.

However, there is another alignment problem. This is not the misalignment of particular AI behaviors with human values, but the misalignment of the effects AI systems can have on the environment, the economy, politics, and other areas of social concern. Some of these concerns are being addressed in attempts to regulate the AI industry.

Regulation

After classifying the types of AI systems, the OECD is primarily interested in the policies that governments can adopt to ensure that rights will not be violated through the improper use of these systems. The European Union has also enacted regulations for AI, and in the official statement of 2024, the purpose of regulation is described as follows:

> The purpose of this Regulation is to improve the functioning of the internal market and promote the uptake of human-centric and trustworthy artificial intelligence (AI), while ensuring a high level of protection of health, safety, fundamental rights enshrined in the Charter, including democracy, the rule of law and environmental protection, against the harmful effects of AI systems in the Union and supporting innovation. [1]

A different approach has very recently been taken by the US government. Instead of legal restrictions and liabilities, tech giants and AI labs are to police themselves, as described in the "White House Accord on Super Intelligence," which was signed by Donald Trump and executives of Google, Anthropic, Meta, OpenAI, xAI, and Nvidia on September 29, 2026 [2].

China does not have a single comprehensive horizontal law like the EU AI Act. Instead, it uses administrative measures issued mainly by the Cyberspace Administration of China (CAC) and other agencies, although a comprehensive national AI law is under legislative research and could be enacted as early as

1. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A02024R1689-20260727
2. https://d3i6fh83elv35t.cloudfront.net/static/2026/09/accord.pdf
   Note that Trump has renamed artificial intelligence [AI] as super intelligence [SI]

2027, "to promote the healthy development and standardized application of generative artificial intelligence, safeguard national security and public interests, and protect the legitimate rights and interests of citizens, legal persons and other organizations."[1] The emphasis is on state-defined ethical norms and social order rather than individual fundamental rights in the European sense.

India lacks a single binding AI Act, instead relying on existing laws plus voluntary guidelines. The main national framework is the India AI Governance Guidelines of November 2025, which sets ethical principles, innovation pathways, and risk-mitigation measures. It explicitly seeks to avoid heavy regulation. The long-discussed Digital India Act has not yet been enacted. It emphasizes human-centric, inclusive, and responsible AI, ethical principles, and risk mitigation. Governance of AI in India is more aspirational than the EU's binding obligations or China's detailed administrative controls. Accountability and ethical considerations are highlighted, with enforcement being lighter and less centralized.

In the measures taken around the globe to advance AI and to avoid the dangers posed by what advanced systems can do, attempts are made to distribute liabilities and responsibilities to individual human beings and corporations. No matter how advanced the AI system, the system itself cannot be held accountable any more than a piece of software. Users, designers, and developers can be held responsible, as well as the corporations that produce and use AI systems. Although there is considerable debate about what guardrails should be set up by governments to avoid risks posed by AI systems, and what steps should be taken to fix the Alignment Problem, existing AI acts and regulations focus overwhelmingly on private-sector actors – providers, deployers, developers, model owners, and users. The legal instruments and regulations do not assign formal legal responsibility to governments for recklessly pursuing what has been described as a race for the development and deployment of AI systems.

Responsibility

Sven Nyholm has addressed the issue of legal responsibility for what automated systems do (Nyholm, 2018). His primary concern is with robotic systems, automated cars, and automated weaponry. Although he does not consider more advanced forms of agentic AI, many of his insights can be extended. The harms caused by such systems have led some ethicists to claim

1. https://www.cac.gov.cn/2023-07/13/c_1690898327029107.htm

that there are responsibility gaps (Matthias, 2004). Responsibility gaps are alleged to occur when none of the humans involved in the use of the system that has caused harm are held to be responsible for the harm done, and yet the system itself cannot be considered responsible. [1] In addition to responsibility gaps due to the use of automated systems, there are also responsibility gaps due to corporate activities. The legal fiction of the corporation as a person may be viewed as a means of filling at least some responsibility gaps that would otherwise result from collective agency. A corporation can be blamed and subjected to legal penalties or restrictions. Since AI systems are not capable of ownership and entering into contractual relationships, they are not considered persons even in the sense of the legal fiction that corporations are.

While discussions of the "responsibility gap" in AI ethics have been dominated by the seminal work of Sven Nyholm – who rightly characterizes these gaps as situations where a system's behavior eludes traditional models of individual culpability – recent work by Herman Veluwenkamp suggests that our conceptual reliance on a singular "responsibility gap" is, in fact, structurally defective. Veluwenkamp argues that we must disaggregate the problem into two distinct, manageable categories: Epistemic Responsibility Gaps (ERG), where the complexity of the system prevents us from knowing who or what caused the action, and Control Misalignments (CM), where the structural delegation of decision-making creates an outcome that no individual actor intended. By adopting Veluwenkamp's distinction, we move away from the futile question of whether AI can be "truly" responsible, and toward a targeted inquiry: is the harm a result of an epistemic failure (which demands greater transparency), or a control failure (which demands better institutional alignment)?

Conclusion: Toward an Attributionist Governance

The search for a singular, essentialist definition of "AI agency" has led the field into a stalemate, oscillating between anthropomorphic fears and reductive functionalism. By shifting the focus from what an AI is to how we attribute agency to it, the paradox of modern AI governance can be resolved. Agency is not a property residing in the code; it is a relational outcome – a socio-technical construct emergent from the interactions between developers, users, regulators, and the system itself.

Our analysis suggests that we must move away from searching for a "soul"

1. This is just a rough introductory characterization of the concept of a *responsibility gap*. For a critical review of several analyses and suggestions, see (Veluwenkamp, 2025).

or "will" in machine learning models and instead cultivate a robust taxonomy of attribution. When we view AI agency as an instrument for distributing responsibility – akin to organizational agency – the "Alignment Problem" loses its existential tenor and gains a practical, manageable focus. We are not aligning "agents" with our values; we are aligning our own complex, often conflicting, regulatory and moral expectations with the systems we instantiate.

The study of artificial intelligence and machine learning systems offers many rewards. It can assist us in our assessments of human agency and actions at individual and social levels. It can prompt us to reconsider issues of divine agency. One of the most valuable lessons comes from acknowledging the contingency of our own human agency. Just as the theological concept of *fāʿil bi-l-tajallī* (agency by self-disclosure) suggests that agency can be expressed through mediated channels without residing in the mediator or in supposed divine intentions, so too should we view AI as a kind of agency without its own intentions: not a divine self-disclosure, but a manifestation of human aims and institutional structures. The goal of our study should not be to discover the mystery of agency, but to ensure that attributions of responsibility function to prevent harm. Those responsible for the actions of autonomous systems include individual developers, owners, designers, and users, as well as the organizations and institutions that make AI deployment possible, not least of which are our governments.

▣ Conflict of Interest

The author declares no competing interests.

References

Anscombe, G. E. M. (1959). *Intention*. Blackwell.

Aristotle. (2014). *The Nicomachean ethics* (C. D. C. Reeve, Trans.). Hackett.

Bagnoli, C. (2021). The objective stance and the boundary problem. *European Journal of Philosophy*, *29*(3), 646–663. https://doi.org/10.1111/ejop.12678

Crook, B. (2026). Risks deriving from the agential profiles of modern AI systems. In V. C. Müller, L. Dung, G. Löhr, & A. Rumana (Eds.), *The philosophy of artificial intelligence: The state of the art* (pp. 67–82). Springer.

Descartes, R. (1972). *Treatise of man*. Harvard University Press.

Ferrero, L. (Ed.). (2022). *The Routledge handbook of the philosophy of agency*. Routledge.

Holzhey, H., & Mudroch, V. (2005). *Historical dictionary of Kant and Kantianism*. Scarecrow Press.

Ibn Sina, H. (1999). *The canon of medicine*. Brigham Young University Press.

Kant, I. (2011). *Groundwork of the metaphysics of morals: A German-English edition* (M. Gregor, Trans.; J. Timmermann, Ed.). Cambridge University Press.

Kelley, M. (2026). A control theory of action. *Australasian Journal of Philosophy*, *104*(3), 689–712. https://doi.org/10.1080/00048402.2025.2521737

Legenhausen, M. (2021). Dimensions of divine freedom. *Journal of Philosophical Theological Research*, *21*(3), 73–91. https://doi.org/10.22091/JPTR.2021.7399.2618

Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology*, *6*(3), 175–183. https://doi.org/10.1007/s10676-004-3422-1

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

OECD. (2026). *The agentic AI landscape and its conceptual foundations*. OECD Publishing.

Russell, S., & Norvig, P. (2020). *Artificial intelligence: A modern approach* (4th US ed.). Pearson.

Strawson, P. F. (2008). *Freedom and resentment and other essays*. Routledge.

Tomasello, M. (2022). *The evolution of agency: Behavioral organization from lizards to humans*. MIT Press.

Veluwenkamp, H. (2025). What responsibility gaps are and what they should be. *Ethics and Information Technology*, *27*, Article 14. https://doi.org/10.1007/s10676-025-09823-8

Wiener, N. (2019). *Cybernetics or control and communication in the animal and the machine* (2nd reissue ed.). MIT Press.

Zhu, L., Lu, Q., Ding, M., Lee, S. U., & Wang, C. (2026). Designing meaningful human oversight in AI. *AI and Ethics*, *6* (3), Article 286. https://doi.org/10.1007/s43681-026-01147-7.

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

Free Agency and Lemos's Indeterministic Weightings Model of Libertarian Free Will

Neil Campbell

Professor Department of Philosophy Wilfrid Laurier University necampbell@wlu.ca

Abstract

I explore John Lemos's recent indeterministic weightings model of libertarian free will and evaluate its promise as a model for human and artificial free agency. First, I disambiguate two different interpretations of the model and clarify which interpretation Lemos prefers. I distinguish between an episodic model and a non-episodic model. According to the first, there are discrete weightings of reasons that occur during deliberation. According to the second, there are no such discrete weightings; instead, reasons are assigned their weights at the end of deliberation just before choosing. I argue that Lemos prefers a non-episodic model that avoids a version of the luck objection, namely, one raised by Haji. I then raise some problems for the non-episodic model. First, I argue following Haji that the model lacks a sufficiently agent-centered account of how an agent assigns weights to her reasons. However, I propose that the model might avoid this problem if it included an appeal to a higher-order setting of priorities by the agent. Despite this possibility, I also argue that without providing a deeper account of the plural voluntary control agents have over the undetermined weighting of reasons, the model is too luck-infected to succeed.

Original Research

Keywords:

artificial agency, free will, indeterminism, deliberation, plural voluntary control

Received: 2026/04/03; **Received in revised form:** 2026/06/12; **Accepted:** 2026/07/15; **Published online:** 2026/07/20

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

Copyright © The Author

Introduction

With rapid developments in Large Language Model artificial intelligence and connectionist systems that learn at unprecedented rates, it is natural to raise questions about the kind of agency of which such systems might one day be capable. Could an artificial intelligence make free and responsible choices, and hence, be a morally responsible agent? As companies rush to incorporate AI into a variety of physical systems that interact with the environment (and with us), from smartphones to autonomous vehicles, the importance of this question becomes increasingly palpable. While it is easy to get caught up in the excitement of the possibilities presented by emerging technology, sober reflection suggests that there is little reason to hope for a clear answer to the above question about artificial agency without first understanding the nature of free and responsible agency in ourselves. It was only by creating models of human intelligence that artificial intelligence became possible, and so it seems that if we want to assess the viability of artificial agency, we need a model of human agency first. If one is a libertarian about human agency, a viable model appears to be a long way off.

Libertarians about free will claim that non-derivatively free choices must be undetermined.1 This is because libertarians believe that to choose freely agents require alternative possibilities, which would be precluded if all choices were determined. The libertarian appeal to indeterminism has taken a variety of recent forms in the literature, leading to several competing models of free agency. To mention just a few, Kane (1996, 1998, 2009a, 2009b, 2014, 2015, 2021, 2023, 2024) describes undetermined efforts of will to choose one alternative over another; Valerian approaches (also sometimes called "modest libertarianism") proposed (though not endorsed) by Mele (1995) and Dennett (1981) appeal to the indeterministic coming to mind of the considerations that figure in deliberation; similarly, Pitman (2012) invokes undetermined acts of the imagination in the generation of conceived alternatives and their potential consequences; while Franklin (2013, 2018) appeals to undetermined basic mental acts over which agents have direct control; and Balaguer (2004, 2012, 2014) describes torn decisions settled by a quantum neurological process. All these models face problems about luck and control. In each case it is claimed by various critics [2] that the presence of an indeterministic process in

1. I use the term "non-derivatively" free choices in contrast to derivatively free choices because of a distinction made by Kane (1998). The former (what Kane calls "self-forming actions") must involve alternative possibilities and be undetermined. The latter can be determined by one's character but are free to the extent that the character from which such choices issue was of one's own free making via previous undetermined (non-derivatively free) choices.

2. The luck-based objections to libertarianism are too numerous to list here, but the following provides a recent sample: (Haji, 2000; van Inwagen, 2000; Mele, 2006; Levy, 2008; Moore, 2021).

deliberation renders the choice too luck-infected to ground free and responsible action. However, in several recent works, John Lemos (2018, 2021, 2023) proposes another libertarian option in the form of his indeterministic weightings model. In the following discussion, I explore Lemos's theory and assess whether his model is any more successful than those that precede it at avoiding the traditional problems of luck and control, and hence, evaluate its potential to serve as a model for human and for artificial agency.

Lemos's Indeterministic Weightings Model: The Basics

I begin with an initial sketch of Lemos's proposal. Lemos's central idea is to locate the indeterminism operative in non-derivative libertarian free choices squarely within the process of deliberation as opposed to putting it at the beginning as Franklin, Pitman, and Valerian theories do, or at the very end as Kane or Balaguer do. Drawing on Nozick (1981) and McCall and Lowe (2005, 2008), Lemos proposes that what is undetermined is the weight that the agent assigns to her reasons during deliberation. Here the somewhat metaphorical notion of "weight" is a matter of the relevance or degree of rational, pragmatic, or other justification a reason provides for one of the alternatives about which the agent is deliberating. That is, when the agent deliberates about whether to choose A or B, one should assume not that each of the reasons for and against A and B has objective weights that the agent tallies up, but that the agent assigns weights to these reasons. Furthermore, the agent's weighting of her reasons is undetermined, meaning that she might have weighted her reasons differently, and if she did, her reasons might have supported a different choice.

To illustrate, consider Lemos's example of Jane deliberating about whether to vacation in Hawaii or Colorado:

> Suppose that as Jane deliberates she assigns in an undetermined way a greater evaluative weight to snorkeling and surfing and lesser weight to mountain views and whitewater rafting, and suppose as well that her choice of the Hawaii vacation is dictated by this assignment of weights. Here we have a vision of her choice as the product of an indeterministic deliberative process which she controls by her assignment of weights to the reasons in favor of each of the options. She ends up choosing the Hawaii vacation because she gave the greater weight to the reasons which support that option, but she could have chosen otherwise had she weighted her reasons differently." (Lemos, 2021, p. 143)

Lemos emphasizes the fact that weighting reasons is something the agent actively and intentionally does. In part, this is to avoid a difficulty he identifies

for Valerian libertarian theories. The problem, Lemos claims, is that on Valerian accounts there is insufficient opportunity for the agent to exercise control over the process such that the resulting choice is both a reflection and a product of the agent's character as opposed to something that merely happens to her. On the Valerian model it is undetermined what considerations come to the agent's mind, but once these arise in an undetermined way in consciousness, they determine the agent's choice. Since the agent does not control what thoughts come to mind or how they lead to choice, it is hard to see how the resulting choice could be an expression of the agent's will. The agent appears to be a mere conduit for these processes. However, things look different if the agent actively weights her reasons during deliberation. Since the weights that are assigned to the reasons that justify a choice are actively and intentionally assigned by the agent, the agent shapes her deliberation in a way that is reflective of her will.

But what of the claim that the weight that the agent assigns to a reason is undetermined? Wouldn't that undermine the idea that the resulting choice is something over which the agent exercises the kind of control required for the above sense of agency and for moral responsibility? If the way the agent weights her reasons is undetermined, isn't the weighting of her reasons and the resulting choice just a product of luck?

To address this, Lemos adopts a strategy employed by Robert Kane. Kane has long argued that an agent is responsible for an undetermined outcome so long as that outcome is the product of the agent's effort. The angry husband who tries to break a glass tabletop is responsible if he succeeds even if, due to some indeterminism in the angular momentum of his arm, he might have failed. Because he accomplished what he was trying to do and succeeded because of his effort, he is responsible for breaking the tabletop even though it was undetermined that he would succeed. As Kane puts it, "It would be a poor excuse for him to say to his wife, 'Chance did it (broke the table), not me'. Though there was a chance he would fail, chance didn't do it, he did" (Kane, 2014, p. 43). Lemos treats the agent's weighting of reasons in the same way:

> … the agent's active assigning of weights to his reasons plays the role of the efforts of will in Kane's view. By assigning a greater value to the reasons for A and lesser weight to the reasons for B, I actively do something which makes me then choose A and not B. On this view, the choice of A doesn't just happen to the agent; rather, just as on Kane's view the agent makes it happen by his actively and intentionally assigning a greater weight to the reasons for A (Lemos, 2018, p. 113).

Lemos also follows Kane in his claim that the agent exercises a compatibilist

variety of control over this intentional activity:

> By meeting … compatibilist requirements while assigning the weights, this ensures that it is the agent who is responsible for assigning the weights; she is not subject to the controlling will of another through force, manipulation, or coercion, and she is not overcome by foreign emotions that do not reflect her higher-order desires (Lemos, 2018, p. 114).

In this schematic form, Lemos's proposal looks like an attractive option for those drawn to libertarianism. First, it accounts for how an agent can have alternative possibilities. Since the way that an agent's reasons are weighted is undetermined, even if the agent's choice follows in a deterministic way from her reasons, she could have chosen otherwise since she might have weighted her reasons in a way that favored the other option. Second, the model does not appear to generate the concerns about irrationality that sometimes plague libertarian accounts of the ability to choose otherwise. Since the agent will (except for cases of weakness of will) choose in accordance with those reasons to which she has assigned greater weight, her choice is guaranteed to fit in a rational and explanatory way with her deliberation. Third, since assigning a weight to a reason is something the agent does intentionally and is something over which she exercises compatibilist-style control, this approach also appears to mitigate the usual concerns about luck. If such a model of freedom of choice holds up to scrutiny, perhaps it could serve to inform attempts to create artificial free agency. At least we would know roughly at what point we need to build indeterminism into an artificial system—assuming such a thing is possible.

To evaluate Lemos's proposal, there are a few details about the position that I need to bring into greater relief. I will focus on three points in the following paragraphs and raise some issues along the way.

When Are Weights Assigned to Reasons?

In *A Pragmatic Approach to Free Will*, Lemos is clear that agents assign weights to reasons prior to choosing. Indeed, this is the primary respect in which his account differs from Nozick's: Nozick holds thatthe weighting of reasons happens at the moment of choice. However, from Lemos's description it is not entirely clear when during deliberation but prior to choice the agent weights her reasons, or how often. One possibility is that the agent assigns a weight to each reason as it occurs to her during the process of deliberation. This would require a series of individual undetermined weightings that follow

each other in sequence as relevant reasons occur to the agent in temporal succession. It might also involve the possibility that an agent assigns a weight to a reason at time t, but later re-weights the same reason in light of new information or further reflection, say at time $t + n$. I will call this the episodic model, since the central idea is that there is a series of episodes of weighting of reasons by the agent throughout the process of deliberation.

At times, it seems as though Lemos intends to motivate an episodic model. In his discussion of objections to libertarianism by Neil Levy (2008, 2011), Lemos describes Jane's deliberation about where to vacation as follows:

> … as she considers the prospects for mountain views and whitewater rafting in Colorado, she may acknowledge that she is from Eastern Tennessee where one can do these things as well, but one cannot surf or snorkel. In light of this, she may give less weight to mountain views and whitewater rafting. Then, she may consider that she fears sharks, and knowing that sharks are present in the waters of Hawaii, she may begin to give less weight to the Hawaii trip. She may look up information on the likelihood of shark attacks and find that they are very rare, and this may lead her to give more weight to the Hawaii trip again (Lemos, 2021, pp. 144–145).

As presented, the example suggests that the agent has a consideration for or against one of the options come to mind, and then assigns that consideration a certain weight, though does so in an indeterministic fashion such that it might have been weighted differently. When Jane thinks about her fear of sharks this is given a certain weight against the Hawaii trip, though when she discovers how rare shark attacks are this is also given a certain weight in favor of the Hawaii trip. Jane's deliberation will, on the episodic model, involve a running tally of the pros and cons for each of the options and the choice will be a product of the net totals, whichever way they lean just prior to choosing.

Before I move on to an alternative interpretation, I want to flag an ambiguity that crops up in the passage just quoted. Under the episodic model the idea is that as reasons for and against each holiday destination occur to Jane, she assigns those reasons a certain weight and those reasons will, depending on the balance of the reasons for and against each trip, dictate which choice she makes. Note, however, that in the passage quoted above Lemos shifts from describing the agent assigning weights to the reasons for or against each option, to simply assigning greater or lesser weight to each trip. Perhaps I am being pedantic, but it does seem to matter which of these two things is happening if we are to be clear about the processes unfolding during deliberation and exactly what part or parts of those processes are

indeterministic. If what is happening is that the agent is assigning greater weight to one of the target choices as opposed to the reasons that support that choice, it becomes more difficult to see how the process of weighting reasons differs from choosing or how Lemos's model differs from Nozick's. Furthermore, the explanatory connection between the agent's reasons and choice loses its resolution if it is the alternative choices that are weighted without first weighting the reasons that justify each choice.

Given that Lemos is elsewhere quite clear that the weightings happen before choosing, I suspect Lemos's intention in the above passage was not to deviate from his earlier claim that agents assign weights to their reasons (as opposed to the choices or options), and so I suspect the ambiguity I have identified is simply the product of Lemos speaking loosely about the process. Setting aside this ambiguity, my first interpretation of Lemos's theory is that agents assign a certain weight to each reason as it occurs to them and then choose in accordance with the option that has the on-balance weightier reasons. I called this the episodic model.

However, in a more recent paper where Lemos (2023) addresses criticisms raised by Haji (2022), he says something that suggests a different interpretation than the episodic model:

> I don't see there being discrete distinct weightings made at different moments during deliberation. The weighting that occurs throughout the deliberation leading up to a choice is a continuous fluid process, and when a decision occurs it is typically made in conformity with where the weights are set at the time of decision (Lemos, 2023, p. 112).

Lemos has a tactical reason for denying that his model involves the discrete weightings described by the episodic model. He wants to avoid a model according to which it is possible to compare exactly similar agents in two contrast worlds at a particular time during deliberation who weight a reason differently at the next moment. Such a possibility would, as Haji argues, generate the intuition that what weight the agent assigns to the reason is a matter of luck, and this is an intuition Lemos understandably wants to avoid. (I will return to the issue of luck later. For now, I just want to clarify Lemos's model.)

If agents do not make discrete weightings of the considerations that occur to them as they come to consciousness during the process of deliberation as the episodic model claims, we are instead to think of the weighting of reasons as a process that continues up to just before the moment of choice. Given the passage just quoted, I do not think that one could even say that a consideration is given an initial weighting that might change over the course of deliberation as the agent revisits some of her reasons in light of new thoughts and ideas. It now seems that

the process of deliberation involves a keeping in mind of what factors are relevant to the choice, but how relevant those factors are (i.e., how much weight they are given) is not decided until just prior to choice. I will call this the non-episodic model since this approach denies that there are discrete episodes of weighting or re-weighting reasons prior to the weighting that immediately precedes choice.

It is worth noting that the non-episodic model fits amicably with Lemos's claim in *A Pragmatic Approach to Free Will* (chapter 5) that he prefers to see deliberation not as a battle between competing efforts as Kane does, but rather as the keeping alive of the options for the agent. When Jane thinks about her fear of sharks this clearly counts against choosing Hawaii as her holiday destination, but until she decides how much it counts it serves only to keep the option of choosing Colorado open for her. Prior to choosing, Jane will settle how much weight her fear of sharks will have (along with the other considerations that occurred to her to that point) and the balance of her reasons will dictate her choice. While this is closer to Nozick's theory than the episodic model considered above, this does maintain a distinction between the agent's weighting reasons and the agent's choosing. Since this non-episodic model appears most recently in Lemos's writings, I will assume that this is his preferred interpretation. Now that we have some clarity on my first question, I want to raise a second.

How Are Weights Assigned to Reasons?

As far as I can see, Lemos does not broach this question directly. However, there is a partial answer in the way he deals with Haji's criticism mentioned above. Recall that Haji raises a version of the luck objection that is directed at the indeterministic weighting of a particular reason. To set the stage for Lemos's response, I need to sketch out Haji's criticism in more detail.

Haji (2022) claims that if the weight an agent assigns to a reason is undetermined, this suggests the following possibility: identical agents Jane and Jane* are deliberating about their holiday destinations in two contrast worlds that are the same up until a certain point, say t_1. At t_2 in world 1, Jane assigns a certain weight to the snorkeling opportunities afforded by Hawaii, while at t_2 in world 2, Jane* assigns less weight to the snorkeling opportunities. Since both agents have the same past and character to that point, Haji claims there is nothing to account for the different weights assigned by the agents, and so it appears to be a matter of luck that Jane and Jane* assigned the weights they did to that reason. If so, their deliberations are too luck-infected to support a free and responsible choice.

Lemos combats this objection in a way that sheds some light on the question of how agents assign weights to reasons. He responds to Haji by

saying that his model is not consistent with the problematic possibility. The reason for this is that Jane and Jane* cannot have the same pasts if they weight their reasons differently:

> For on my view, if the options are weighted differently whether at the final stage of deliberation just prior to choice or at some intermediate stage of deliberation, then there must have been some differences in the paths that the indeterministic deliberative processes of Jane and Jane* went during deliberation—different ideas or reasons for the options must have come to mind and/or how they get weighted must have been different, etc. (Lemos, 2023, pp. 109–110).

Here Lemos incorporates a Valerian element into his own theory by claiming that Jane and Jane* cannot be the same at some intermediate stage of deliberation because if the way they weight a particular reason diverges, this can only be because different thoughts indeterministically occurred to each of them prior to that point. So, Lemos assumes not only that it is undetermined how an agent weights a reason in deliberation, but that some of the thoughts that occur to an agent prior to any weighting are also undetermined. This is clearly a Valerian addition since Jane and Jane* can only differ in their flow of thoughts if we assume it is undetermined what thoughts occur to them. Most significantly, this Valerian element seems to provide the explanation for why an agent assigns a certain weight to a reason. Since Jane and Jane* assign different weights to the same reason because different thoughts occurred to each of them before that point, it seems the prior thoughts are, at least in part, what explain their weightings.

By constructing his example in the way he does, Haji clearly interprets Lemos's theory as an episodic model. The claim that Jane and Jane* could be identical in a cross-world comparison and yet weight a reason differently is only possible on such an interpretation. I have suggested, however, that Lemos actually prefers a non-episodic model, according to which Jane and Jane* will weight a reason only at the conclusion of deliberation, and given that it is undetermined what thoughts occur to each of them prior to that point, it is wildly implausible to suggest that Jane and Jane* could have identical histories by the time they assign weights to their reasons. Given this, Lemos is surely correct that Haji's criticism misfires. However, Haji raises another important point about this way of avoiding the luck objection that takes us back to my question about how agents assign weights to their reasons.

Haji (2023) has complained that Lemos's way of avoiding the luck objection described above is a "non-action centered" approach that is not congenial to Lemos's own emphasis on the idea that by weighting reasons actively and intentionally, the resulting choice reflects the agent's will. I am

sympathetic with Haji on this point. After all, the reason Lemos does not embrace a wholly Valerian approach to libertarian free will is, as I observed earlier, that he thinks it does not give the agent enough control over the weighting (and choice) to see these as expressions of the agent's will or character. If the only thing to which Lemos can appeal to explain why Jane and Jane* make their respective weightings is the undetermined thoughts that came to their minds, this appears to erode the kind of control Lemos wants agents to have over the process of deliberation.[1] So, while I think Haji's criticism targets an episodic rather than non-episodic model and that Lemos advocates the latter rather than the former, there is still a problem about how to explain the agent's weighting of a reason. Appealing to Valerian non-actional processes, such as what different thoughts popped into Jane and Jane*'s minds, threatens to render the non-episodic version of Lemos's theory vulnerable to the very criticism he raises against Valerian theories.

Of course, it may be that the Valerian element described above is only part of the story, and perhaps not the main part. After all, Lemos also emphasizes the claim that agents exercise a compatibilist form of control over the process of weighting reasons. I will return to this point in the next section. First, I would like to raise a possibility that might be helpful to Lemos's non-episodic model. As Haji points out, what is needed in order to provide an explanation for how an agent weights a reason in a way that helps us to see the weighting (and the resulting choice) as an expression of the agent's will is something more agent-centered than the thoughts or considerations that happened to occur, indeterministically, to her. I suggest that Lemos might look to Patricia Greenspan for help here.

Greenspan (2012) has also proposed an indeterministic weightings model of libertarian free will and appeals to an agential element that might help Lemos avoid the identified problem. According to Greenspan's model of "active deliberation," the way an agent weights her reasons is a matter of a hierarchical setting of priorities. Her example involves trying to decide whether to vacation in Rome or on the Italian Riviera. Let us assume that she has reasons that rationally support each option equally. Greenspan claims that she can make one set of reasons prevail by prioritizing her values. If she prioritizes relaxation over cultural enrichment, she will weight the reasons for choosing the Riviera

1. A further problem with Lemos's response to Haji's criticism is that he arguably plays fast and loose with the very meaning of "indeterminism" itself. As Kane has pointed out, "If my choice is really undetermined, that means I could have made a different choice given exactly the same past right up to the moment when I did choose. That is what indeterminism and probability mean: exactly the same past, different possible outcomes" (Kane, 2002, 226). So, if Lemos claims that how an agent assigns weight to a reason is undetermined, that would seem to imply the possibility he denies but that Haji insists on.

over her reasons for going to Rome and so will choose to vacation on the Riviera. Alternatively, if she prioritizes cultural enrichment over relaxation, her reasons for going to Rome will be assigned more weight and that is what she will choose. Greenspan proposes that the setting of priorities by the agent is undetermined, meaning that which of the agent's values are prioritized is arbitrarily decided by the agent. Since Greenspan thinks there is nothing rationally deficient or puzzling about an agent arbitrarily deciding to prioritize one of her values over another, she proposes that her model allows for the kind of "rational swerves" required for the libertarian ability to choose otherwise.

It seems Lemos's model might benefit from an appeal to something similar to ensure that he has a sufficiently agent-centered account of how agents weight their reasons. To see how, let us reconsider the example appealed to in his disagreement with Haji. Jane and Jane* are two exactly similar agents in two contrast worlds. They begin to deliberate about whether to vacation in Hawaii or Colorado. While it may be that different ideas, images, and considerations indeterministically rise to consciousness in their respective minds during this process, perhaps it is also possible that Jane indeterministically prioritizes the value of excitement and so weights white water rafting very heavily, while Jane* prioritizes the value of peace and tranquility and so weights the activity of snorkeling with colorful species of fish very heavily instead. Since Jane has prioritized excitement, she weights the reasons for vacationing in Colorado more heavily than her reasons for vacationing in Hawaii and chooses the Colorado trip. Since Jane* prioritizes tranquility, she weights the reasons for vacationing in Hawaii more heavily and chooses to go there. As with Lemos's original model, we preserve a rational and explanatory connection between the agent's reasons and her choices, but now we can say more about why Jane and Jane* weighted their reasons differently and went on to make different choices. Furthermore, by appealing to the agent's prioritization of one of her values, Lemos can appeal to something about the agent that accounts for the difference between Jane and Jane*. Since their weightings differ because of what they decide to value more, we can see their weightings and subsequent choices as expressions of their character—of who they are as persons and what they value—rather than as the product of processes over which they have no control. Although I doubt that this friendly amendment to Lemos's theory would render it immune to worries about luck, it might provide part of a more satisfying agent-centered explanation for the assigning of weights to reasons.

How Is the Indeterminism in Deliberation and Choice Being Conceptualized?

Robert Kane has offered what is probably the most detailed and elaborate

account not only of the role that indeterminism might play in libertarian free will, but also of how we should conceptualize this indeterminism in the context of deliberation and choice. A weakness of his account is that it depends on a conjecture—albeit a scientifically informed one—about the magnification of small-scale indeterminacies at the synaptic level to the higher-level information processing of the brain, and this conjecture might be mistaken. One of the great strengths of Kane's approach, however, is his proposal that we conceptualize the indeterminism at work in deliberation and choice as a hindrance or obstacle to the attainment of a goal.

As far as I can tell, Lemos is agnostic about the underlying indeterministic physical processes, preferring instead to motivate the belief in libertarianism on pragmatic grounds. But on the question of how we should conceptualize the role of indeterminism in deliberation and choice, he follows Kane's lead:

> Recall Kane's example of the angry husband that makes the effort to break the glass table and succeeds in breaking it. Even if the table's breaking is not a determined consequence of his making the effort, he is still in sufficient control over what happens to make him fully responsible for it. So, too if a choice is the undetermined result of agential efforts or the weightings of reasons made in deliberation, we can be fully responsible for the choice (Lemos, 2025, p. 982).

This passage suggests that Lemos, like Kane, thinks of indeterminism as a hindrance or obstacle to be overcome via a purposeful activity of the agent. Unlike Kane, though (at least until the recent shift in Kane's thinking on this issue; see Kane, 2021, 2023, 2024), Lemos prefers to avoid the idea that an agent makes dual efforts to assign competing weights to a reason. This is because Lemos (2011) has long found the dual efforts component of Kane's theory problematic. Instead, he appeals to the variety of compatibilist control that Kane claims we have over efforts of will. Lemos proposes that this variety of control is sufficient to ensure that even though how much weight an agent assigns to a reason in deliberation is undetermined, the agent has enough control over the weighting for us to agree that it is the agent (not luck) that assigns that weight and, hence, is responsible for the resulting choice.

But here there seems to be a problem. Although the angry husband might succeed or fail in his attempt, he is not trying to fail, so he can only fail by accident. Since a non-derivatively free choice is the product of an indeterministic process and that process can lead to either of two possible outcomes, Kane's use of examples like the angry husband can only work if the agent is simultaneously trying to secure both results. This is why Kane insisted for years that agents engage in dual efforts of will: it ensures that whichever

way the choice goes, the agent is responsible for it. The challenge for Lemos appears to be as follows: how can the compatibilist style of control over the weighting of a reason convince us that it is the agent and not luck that gives that reason the weight it does since the reason might have been weighted differently? In the absence of something like simultaneous efforts to assign distinct weights to the reason, the answer is not obvious.

A slightly different way to frame this important issue is in terms of the concept of plural voluntary control (or "PVC"). Since the agent might assign different weights to a reason, she needs to have PVC over the weighting if we are to avoid the intuition that how she weights a reason is a product of luck. Kane at one time understood PVC as follows:

> [It is] the power to make either choice be or not be at the time, voluntarily (without being coerced or compelled in doing so), intentionally (on purpose, rather than merely accidentally or by mistake or unintentionally), and for the reasons one had for making that choice rather than the competing choice (Kane, 2019, p. 150).

Lemos also embraces PVC:

> To act with PVC is to do something, like making the choice of A instead of B, when in the moments leading up to the choice you had the ability to do otherwise such that you could have chosen A or B and been in control of either choice. But we can only exert PVC over an undetermined choice by having compatibilist control over events that transpire in the deliberation process which leads to the choice—whether the efforts of will (Kane's view) or the assigning of weights to reasons (my view) (Lemos, 2023, p. 106).

Under Kane's dual efforts model, the voluntariness condition of PVC is satisfied by compatibilist standards of control, but these standards need to apply to both efforts simultaneously. In the absence of dual efforts, the agent can only choose otherwise by accident. What appears to be missing in Lemos's treatment of PVC is an account of how the agent can exercise compatibilist-style control over (at least) two different weightings without appealing to something like Kanean dual efforts.

Lemos thinks that such a move is unnecessary because of his claim that which weight is assigned will be a product of the thoughts that occurred, indeterministically, to the agent prior to the weighting. If we compare Jane and Jane* in their respective worlds and accept the idea that it is undetermined what thoughts come to their respective states of consciousness, it might appear to be enough for them to take ownership of their respective weightings if they

weight the reasons in question voluntarily (i.e., they exercise compatibilist control over their weightings). However, we already saw that appealing to this non-actional Valerian element is insufficient since it does not give the agent enough control to be the author of the weighting (or the resulting choice) since the agent has no control over what thoughts come to mind. Lemos needs plural voluntary control if the assignment of weight to a reason can go more than one way rationally and intentionally. I am at a loss to see how compatibilist freedom from "the controlling will of another through force, manipulation, or coercion, and … foreign emotions that do not reflect [the agent's] higher-order desires" alone can secure plural voluntary control over the weighting of a reason.

In a discussion of Kane's most recent model of libertarian free will, where he moves away from simultaneous dual efforts of will toward sequential singular efforts, Dwayne Moore (2025) draws a helpful distinction between different ways of thinking about PVC that can shed some additional light on this issue. He distinguishes between synchronous PVC where "the agent had the power and opportunity to make either choice be or not be at the time …" (Kane, 2019, p. 150) and asynchronous PVC, "according to which agents must merely have the power and opportunity to make either choice be in the course of deliberation" (Moore, 2025, p. 9). Kane's dual efforts model embraced synchronous PVC, but his most recent model adopts asynchronous PVC instead. The fundamental difference between them is the way that the ability to choose otherwise is understood against a temporal component. To have synchronous PVC I must be able to choose and choose otherwise right now, with the same past and deliberation. For asynchronous PVC I do not need to be able to choose and choose otherwise right now, at the moment of choice, but it must have been possible, due to indeterministic processes that happened during deliberation, for my deliberation to have turned out differently than it did, and if it had run a different course, I would have made a different choice. Lemos's non-episodic indeterministic weightings model clearly embraces asynchronous PVC since the agent's choice will be dictated by the weights she assigns to her reasons, but those weights (prior to choice) might have been different because they are assigned by the agent indeterministically. But for Lemos's non-episodic indeterministic weightings model to secure asynchronous PVC over the choice that the agent makes, the agent also needs to have PVC over the weighting of her reasons. I have proposed that the agent can only have synchronous PVC over how she weights a reason if Lemos appeals to something like dual efforts to assign two different weights simultaneously—something Lemos wants to avoid.

Lemos's strategy appears to be to rely on asynchronous PVC over the weighting of reasons. As he points out, Jane and Jane* will only diverge in the

weight they assign to a particular reason (or to the overall balance of their reasons) if different thoughts or considerations occurred, indeterministically, to each of them. But as Lemos himself observes in the quotation above, we "can only exert PVC over an undetermined choice by having compatibilist control over events that transpire in the deliberation process." By parity of reasoning, it seems we can only have asynchronous PVC over the weighting of a reason by having compatibilist control over indeterministic processes that precede the assignment of weights to reasons. The problem is that it is reasonably clear that we lack any control (compatibilist or otherwise) over which thoughts occur, indeterministically, to us. Unless Lemos can secure a stronger form of plural voluntary control over the indeterministic weighting by an agent of her reasons, there will always be a danger that the agent's choice will appear to be too luck-infected to be a free and responsible choice.

Conclusion

Although there is much to recommend Lemos's theory, I am not optimistic that an indeterministic weightings model of free agency could work either for us or for an artificial system. I have identified some problems such a model would need to overcome. First, it would require a more agent-centered account of how agents assign weights to the reasons that occur to them in deliberation. I proposed that an appeal to a higher-order setting of priorities might provide some help to this problem. However, even with such an amendment, a daunting difficulty remains. Given the largely Kanean way that Lemos conceptualizes the indeterminism involved in the assigning of weights to reasons, it is hard to see how agents could have the kind of plural voluntary control over this process that is needed to ensure that however an agent weights a reason we can be confident that she (not luck) is responsible for that weighting.

▣ Conflict of Interest

The author declares no competing interests.

References

Balaguer, M. (2004). A coherent, naturalistic, and plausible formulation of libertarian free will. *Noûs*, *38*(3), 379–406. https://doi.org/10.1111/j.0029-4624.2004.00475.x

Balaguer, M. (2012). *Free will as an open scientific problem*. MIT Press.

Balaguer, M. (2014). *Free will*. MIT Press.

Dennett, D. C. (1981). *Brainstorms: Philosophical essays on mind and psychology*. MIT Press.

Franklin, C. E. (2013). How should libertarians conceive of the location and role of indeterminism? *Philosophical Explorations: An International Journal for the Philosophy of Mind and Action*, *16*(1), 44–58. https://doi.org/10.1080/13869795.2013.723036

Franklin, C. E. (2018). *A minimal libertarianism: Free will and the promise of reduction*. Oxford University Press.

Greenspan, P. (2012). Free will and rational coherency. *Nous-Supplement: Philosophical Issues*, *22*, 185–200. https://doi.org/10.1111/j.1533-6077.2012.00224.x

Haji, I. (2000). Indeterminism, explanation, and luck. *Journal of Ethics: An International Philosophical Review*, *4*(3), 211–235. https://doi.org/10.1023/A:1009817905421

Haji, I. (2022). Libertarianism and luck. *Journal of Philosophical Theological Research*, *24*(3), 115–134. https://doi.org/10.22091/jptr.2022.8297.2723

Haji, I. (2023). Further reflections on Lemos's indeterministic weightings model of libertarian free action. *Journal of Philosophical Theological Research*, *25*(3), 119–130. https://doi.org/10.22091/jptr.2023.9722.2929

Kane, R. (1996). Freedom, responsibility, and will-setting. *Philosophical Topics*, *24*(2), 67–90. https://doi.org/10.5840/philtopics199624210

Kane, R. (1998). *The significance of free will*. Oxford University Press.

Kane, R. (2002). Free will: New directions for an ancient problem. In R. Kane (Ed.), *Free will* (pp. 222–248). Blackwell.

Kane, R. (2009a). Free will and the dialectic of selfhood: Can one make sense of a traditional free will requiring ultimate responsibility? *Ideas y Valores*, *58*(141), 25–43.

Kane, R. (2009b). Libertarianism. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, *144*(1), 35–44. https://doi.org/10.1007/s11098-009-9365-y

Kane, R. (2014). Acting 'of one's own free will': Modern reflections on an ancient philosophical problem. *Proceedings of the Aristotelian Society)*, *114*(1pt1), 35–55. https://doi.org/10.1111/j.1467-9264.2014.00363.x

Kane, R. (2015). On the role of indeterminism in libertarian free will. *Philosophical Explorations*, *19*(1), 1–15. https://doi.org/10.1080/13869795.2016.1085594

Kane, R. (2019). The complex tapestry of free will: Striving will, indeterminism and volitional streams. *Synthese*, *196*(1), 145–160. https://doi.org/10.1007/s11229-016-1046-8

Kane, R. (2021). Making sense of a free will that is incompatible with determinism: A fourth way forward. *Journal of Philosophical Theological Research*, *23*(89), 5–28. https://doi.org/10.22091/jptr.2021.7247.2595

Kane, R. (2023). Free will, complexity, dynamical systems, and all that jazz. *Harvard Review of Philosophy*, *30*, 1–22. https://doi.org/10.5840/harvardreview20245856

Kane, R. (2024). *The complex tapestry of free will: A philosophical odyssey*. Oxford University Press.

Lemos, J. (2011). Wanting, willing, trying and Kane's theory of free will. *Dialectica*, *65*(1), 31–48. https://doi.org/10.1111/j.1746-8361.2010.01258.x

Lemos, J. (2018). *A pragmatic approach to libertarian free will* (Vol. 107). Routledge.

Lemos, J. (2021). The indeterministic weightings model of libertarian free will. *Journal of Philosophical Theological Research*, *23*(3), 137–156. https://doi.org/10.22091/jptr.2021.7187.2582

Lemos, J. (2023). Haji and the indeterministic weightings model of libertarian free will. *Journal of Philosophical Theological Research*, *25*(3), 101–118. https://doi.org/10.22091/jptr.2023.9648.2920

Lemos, J. (2025). Replies to Caruso, Elzein, Pereboom, Sekatskaya, and Smilansky. *Philosophia (Ramat Gan)*, *53*(3), 973–993. https://doi.org/10.1007/s11406-025-00904-y

Levy, N. (2008). Bad luck once again. *Philosophy and Phenomenological Research*, *77*(3), 749–754. https://doi.org/10.1111/j.1933-1592.2008.00219.x

Levy, N. (2011). *Hard luck: How luck undermines free will and moral responsibility*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199601387.001.0001

McCall, S., & Lowe, E. J. (2005). Indeterminist free will. *Philosophy and Phenomenological Research*, *70*(3), 681–690. https://doi.org/10.1111/j.1933-1592.2005.tb00420.x

McCall, S., & Lowe, E. J. (2008). The determinists have run out of luck—For a good reason. *Philosophy and Phenomenological Research*, *77*(3), 745–748. https://doi.org/10.1111/j.1933-1592.2008.00218.x

Mele, A. R. (1995). *Autonomous agents: From self-control to autonomy*. Oxford University Press.

Mele, A. R. (2006). *Free will and luck*. Oxford University Press.

Moore, D. (2021). Libertarian free will and the physical indeterminism luck objection. *Philosophia*, *50*(1), 159–182. https://doi.org/10.1007/s11406-021-00370-2

Moore, D. (2025). Agent-causal, event-causal, libertarian free will. *Inquiry*, 1–23. https://doi.org/10.1080/0020174X.2025.2576137

Nozick, R. (1981). *Philosophical explanations*. Harvard University Press.

Pitman, M. M. (2012). Freedom, indeterminism and imagination. *South African Journal of Philosophy*, *31*(2), 369–383. https://doi.org/10.1080/02580136.2012.10751782

Van Inwagen, P. (2000), The Eighth Philosophical Perspectives Lecture: Free Will Remains a Mystery. Noûs, 34: 1-19. https://doi.org/10.1111/0029-4624.34.s14.1

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

Artificial Intelligence, Consciousness, Transcendence: A Thoughtful Overview between Present and Future

Paolo Di Sia

Professor, School of Engineering and School of Medicine, University of Padova, Padova, Italy, paolo.disia@unipd.it

Professor, School of Medicine, Engineering for Innovation Medicine, University of Verona, Verona, Italy, paolo.disia@univr.it

Original Research

In recent decades, artificial intelligence (AI) has gradually moved beyond the purely technical and computer-based realm and is becoming a central topic of cultural, philosophical, ethical, and even theological debate. Having started as an engineering project to create machines capable of performing complex tasks and calculations, it is now becoming a challenge with fundamental questions connected to thought, understanding, and consciousness. Starting from a systematic introduction, the paper thoughtfully analyzes the distinction between strong and weak AI, including the main philosophical criticisms, the ethical and social implications of AI development, the problem of artificial consciousness, the Vatican's thinking on the matter, the future perspectives for AI, and a final critical synthesis.

Keywords:

artificial intelligence, consciousness, transcendence, theology, human being, science and technology, culture and society, human rights.

Received: 2026/04/23; **Received in revised form:** 2026/06/07; **Accepted:** 2026/07/22; **Published online:** 2026/07/25

Di Sia, P. (2026). Artificial Intelligence, Consciousness, Transcendence: a thoughtful Overview between Present and Future. *Journal of Philosophical Theological Research*, (Special Issue on AI, Consciousness, and Human Agency), 28(1), 109-126. https://doi.org/10.22091/jptr.2026.17615.3698

Copyright © The Author

Introduction

All human cultures have experienced innovations throughout their history, but never has anything as new as AI caused such a shock to human consciousness. Its fields of application are multiplying in all daily activities, and it is becoming difficult to preserve our privacy and freedom.

Our era is characterized by the replacement of human intelligence with electronic machines, which cannot escape the needs that manufacturers have imposed on them, and we do not clearly know what these manufacturers want.

Considering man's anthropological need to rely on an external entity capable of providing absolute certainty, we can consider AI as the final step in a long process of delegation, which begins with mythological deities and, passing through religious infallibility, arrives at the presumed computational objectivity.

Considering the words "Artificial Intelligence," we ask ourselves: can an artifice be intelligent in itself? An artifice is a product, the result of a "production," and is not an "action." Ancient philosophy similarly distinguishes the word "*ratio*" from "*intellectus*." Reason has a function of calculation, of comparing quantities, and does not go beyond that task. Intelligence, on the other hand, is creative; it does not know the limits of reason, sees the specifically human qualities of a work or of an action, goes beyond the immediate, and aims at a further fulfillment and at a value.

In recent decades, AI has progressively moved beyond the purely technical and computer-based sphere to become a major topic of philosophical, cultural, and ethical debate. Born as an engineering project aimed at building machines capable of performing complex tasks, AI has now become a challenge for contemporary thought, raising questions such as: is it (and will it be) still possible to distinguish humans from machines? Can a machine really think? Can it understand? Can it be conscious?

The most advanced linguistic, logical, and creative capabilities of AI challenge the very concept of mind and raise profound questions about the nature of intelligence, necessitating a decisive rethinking of traditional notions of consciousness, intentionality, and understanding. The contrast between strong and weak AI is a crucial theoretical issue regarding thinking, understanding, and consciousness. The question is not purely technical, but involves epistemological, ontological, and ethical aspects.

Intelligence is a common-sense concept, and no particular discipline can claim a monopoly on it; dozens of different definitions have been proposed to date, mostly disputed between psychology, educational sciences, and cognitive sciences.

We can define intelligence as the ability to reason, plan, solve problems, develop abstract thought, grasp complex ideas, learn quickly, and draw from experience. It manifests itself primarily in calculation, reasoning, perception of relationships and analogies, rapid learning, memorizing and retrieving information, speaking fluently, classifying, generalizing, and adapting to new or unfamiliar situations.

The paper is structured as follows:

The distinction between strong and weak AI is considered, starting from a theoretical perspective and comparing the distinct positions of authors who have worked on the subject (§ 2).

The following is an in-depth analysis of the main philosophical criticisms leveled at strong AI, considering in particular experiments, problems, and phenomenological objections. The topic concerns the so-called "artificial mind" (§ 3).

AI has led to macroscopic ethical and social implications, including a sort of redefinition of the human being, the delicate question of possible "machine rights," and the impact of automation on the world of work and on the question of inequality (§ 4).

Another sensitive and timely issue concerns the problem of so-called "artificial consciousness," that is, whether a machine can "feel" in a similar way to human feeling, evaluating neuroscientific and computational contributions. This leads to reflection on the future prospects of AI, with the resulting potential ethical, social, and philosophical implications, from the decision-making autonomy of machines to algorithmic justice, and from the protection of individual freedom to that of personal dignity (§ 5).

The Vatican's pronouncements on the matter are considered, beginning with the first discussions on AI dating back to around 2018 with Pope Francis, and ending with the latest reflections on the matter by Pope Leo XIV (§ 6).

The cultural and anthropological transformations triggered by AI constitute an important element for reflection, particularly highlighting hybrid models and the potential emergence of an "artificial consciousness." The paper attempts a critical synthesis that highlights the limits of AI and provides elements of reflection for the future in a transdisciplinary approach.

This is a complex debate that concerns the present and the near future, centered on the human being as a thinking being. Fascinating and disturbing scenarios emerge, such as the distinction between simulation and reality, behavior and understanding, and functional intelligence and phenomenal consciousness. The conclusions follow (§ 7).

Reflecting on Weak and Strong AI

Around 1980, an essential distinction was introduced between strong and weak AI; according to weak AI, computers can simulate the mind, whereas strong AI asserts that a suitably structured program is a mind (Bory et al., 2025; Sheikh et al., 2023). Various ideas have been proposed as critiques of strong AI, including the well-known "Chinese Room" thought experiment, which addresses the lack of understanding associated with mere symbol manipulation (Searle, 1980).

This perspective undermines the claim of strong AI to explain consciousness and intentionality through pure computation; computers do not possess mental states, as they mimic intelligence but do not truly think. The human mind is not a program, and consciousness cannot be reduced to a computational process.

The experiment has faced criticism from one side (Dennett, 1992), and on the other hand there have been intermediate attempts (Chalmers, 1997), proposing a functional strong AI that, despite lacking phenomenal consciousness, possesses structures analogous to human ones.

The two positions, which originated in different fields (one more technical and engineering-oriented, the other more philosophical), have influenced each other, with boundaries that are not entirely clear-cut. If a machine produces coherent responses in a given language, this does not imply that it actually understands that language.

This distinction between "simulation" and "understanding" leads to the broader separation between "function" and "consciousness"; a machine can mimic intelligent behavior but has no lived experience, no subjectivity. It is a difference of an ontological nature.

Accepting the strong AI thesis means accepting that consciousness, understanding, and intentionality – that is, what makes us human – can be reduced to a set of computational operations. Conversely, if we regard subjectivity as irreplaceable, then strong AI remains a simulation; however accurate it may be, it cannot become a potential moral subject.

The contrast between strong and weak AI represents a divergence of an epistemological and anthropological nature; weak AI fits within the logic of technology, whereas strong AI challenges the fundamental contemporary conceptions of mind, consciousness, and subjectivity.

The debate on AI is therefore not merely an engineering issue; it is certainly fertile ground for philosophy, as it touches upon the shared concept of humanity (Di Sia, 2021, 2026; Eysenck & Eysenck, 2022).

The "Artificial Mind": Computation or Understanding?

Computers operate on symbols according to formal rules, but do not attribute meaning to what they process; the simulation of linguistic behavior does not inevitably imply true understanding, contrary to the view that "the brain is a computer (hardware) and the mind is a program (software)."

The problematic stance of strong AI is that it conflates the external appearance with the actual conscious (internal) experience, despite facing criticisms such as Dennett's "systems hypothesis" (Dennett, 1987).

Chalmers (1995) revisited this debate by introducing a distinction between "easy problems" and the "hard problem" of consciousness. The easy problems concern cognitive functions such as language, learning, perception, and attention, which can be scientifically addressed and even replicated in a machine. The hard problem, however, is different, because it concerns the fact that certain experiences are accompanied by conscious states, by "subjective sensations" (Chalmers, 1995).

This concerns the concept known as the "philosophical zombie"; a machine could behave exactly like a human being without possessing any consciousness, a reality identical to that of a human in terms of behavior, yet devoid of awareness.

Dreyfus (1992) has also leveled criticism at strong AI, arguing that human thought cannot be described in exclusively symbolic and logical-deductive terms; human intelligence is not based on the pure and abstract manipulation of symbols, but rather involves a bodily engagement situated in the world.

Distinctions between access consciousness and phenomenal consciousness have been proposed (Block, 2011) along with neurophilosophical models of consciousness (Metzinger, 2010), highlighting the role of the body and the sense of self. However, the question of intentionality remains unresolved; a machine does not possess it in the full sense, even though it can simulate it.

The differences between pairings such as simulation and understanding, behavior and consciousness, and symbol and meaning raise major questions that concern not only the philosophy of mind but also ontology, phenomenology, and ethics, sparking heated debates around the idea that it is possible to "build" a mind, in the human sense of understanding and feeling.

On the Ethical and Social Implications of AI

Strong AI challenges humanity's deepest certainties, suggesting a world in which the distinction between the thinker and that which simulates thought is

becoming increasingly shaded (Bostrom, 2014). Accepting, even if only theoretically, the idea that AI could become conscious leads to the inevitable issue of machine rights, raising questions of liability should a machine violate regulations or cause damage.

Increasing automation carries the risk of causing structural unemployment across many professions, particularly those involving repetitive, low-skill tasks; entire sectors have already undergone transformation and could be completely replaced by artificial agents. This prompts reflection on the need for an adequate and shared regulation of technological development to prevent the concentration of power and the emergence of new forms of economic domination.

These are ethical challenges linked to the question also known as the "trolley problem" in an algorithmic version (Foot, 2003). Monitoring and surveillance, increasingly employed by companies and governments, are also expanding without adequate common regulation, raising serious concerns regarding privacy, individual freedom, and democracy.

One of the risks is that AI could effectively become an instrument of repression, limiting or eliminating fundamental rights. This represents a new form of power, not based on overt force, but on the "algorithmic control" of individuals, which predicts and influences human behavior through data (Di Sia, 2025a; Zuboff, 2019).

The inevitable ethical and social implications of AI underscore the fact that technology can never be considered neutral; every innovation is linked to a set of possibilities, risks, and responsibilities, with significant impacts on the way human beings live, work, and relate to one another.

The future of AI depends on the choices we make today regarding areas of application, limits to be set, and values to be upheld. Philosophy has a role to play in this debate, fostering critical thinking and serving as a guide toward the common good.

The ethical and social implications of AI are constantly evolving, and technology alone cannot drive such change; what is needed is serious philosophical reflection, shared governance, and a public ethic capable of steering AI development toward human-centered, socially responsible, and sustainable goals. Only by following this path will it be possible to transform AI from a potential threat into a valuable resource serving the dignity and freedom of everyone.

Going Beyond AI: "Artificial Consciousness"

The concept of consciousness refers to subjective experiences, qualitative

mental states, and an orientation toward personal identity, aspects absent from a purely functional or computational environment, even those associated with strong AI, which aims for a complete replication of mental faculties, including subjectivity.

Attempts have been made to explain consciousness in mechanical-biological terms through neuroscientific models. Among the best-known, there is Baars's "Global Workspace Theory" (Baars, 2005), taken up and further developed by Dehaene, according to whom consciousness would emerge through a form of global information integration across different brain areas.

Also Tononi et al. (2016), with the "Integrated Information Theory," quantifies consciousness in terms of the degree of information integration and differentiation, regarding as conscious any physical system that reaches a certain critical value of this integration.

The central problem can therefore be reformulated as follows: is a simulation of a mind "really" a mind? Drawing also on Searle's arguments, the answer is negative: an algorithm that simulates emotions may not actually feel anything, as simulating them is not equivalent to having them. For many authors, this depends on the physical substrate: a computer, with hardware made up of electrical circuits and electronic components, cannot generate consciousness, because it lacks the necessary biological conditions for this to occur.

Another line of criticism directed at strong AI comes from phenomenology and the so-called "embodied approaches," supported by thinkers such as Merleau-Ponty, Dreyfus, and Gallagher, which hold that it is impossible to "disembody" consciousness; rather, consciousness is found in the body's interaction with the world, within a localized and reflective context (Merleau-Ponty, 2013).

AI, by contrast, relies on abstraction and symbolic manipulation, ignoring the fundamental role of the body; a machine cannot undergo experiences or feel emotions, desires, fears, anguish, love, or loneliness. It lacks, in short, the embodied essence of conscious experience. These are questions that currently lack a definitive answer, yet they highlight the profound gap between artificial reality and subjective presence. It is as if the "divine spark" that makes us human is missing.

Chalmers has argued that no purely physicalist or computational theory can fully explain the emergence of consciousness, and proposed alternative hypotheses, such as informational panpsychism, that view consciousness as a fundamental property at a universal level. This perspective, which regards consciousness not as an output but as a primordial reality, is also supported by others, such as Faggin (2022, 2024), and is a fundamental element of recent unified holistic theories (Di Sia, 2023, 2024, 2025b).

The relationship between consciousness and AI is one of the most fascinating, unresolved, and disturbing problems in contemporary philosophy; on one hand, there are those who hypothesize that consciousness can be replicated through complex information-processing structures, while on the other, there is a firm insistence on the unrepeatability of subjective human experience and on the irreducibility of the self.

The possibility of a conscious artificial mind involves not only ontological and epistemological questions but also ethical, existential, and cultural ones. To inquire into artificial consciousness is, ultimately, to question our very essence as human beings, as well as what we might become and what it is legitimate not to lose.

AI, Religion, and the Vatican's Stance on the Matter

Throughout human history, humanity has often sought absolute points of reference to which to delegate the weight of the most difficult choices, the explanation of the deepest mysteries, and guidance during the darkest times. For centuries, this role was fulfilled by religion in its many forms; today, in the era of technological post-humanism, AI has emerged as a new object of faith.

Since ancient times, religion has offered ultimate and infallible answers to humanity's fundamental questions. In Archaic Greece, the Olympian deities governed the cosmos and morality; in the Judeo-Christian tradition, God is infallible by definition, and is omnipotent, omniscient, and omnipresent. The religious person accepts their own fallibility and seeks refuge in a higher truth that is not subject to error.

With the Enlightenment and the advent of modern science, the monopoly on absolute truth began to waver, with the emerging idea that human reason, if properly employed, could replace divine revelation. Thus, the human need for certainty did not disappear but gradually shifted toward science, redirecting the need for infallibility.

In today's hyper-connected, algorithm-driven society, AI is increasingly perceived as possessing a higher rationality, free from prejudices, and acting as a sort of "oracle machine" capable of making better decisions than humans in every sphere.

The infallibility attributed to AI derives from two widely held assumptions: the belief that data is neutral and the conviction that computation is synonymous with truth. However, as highlighted by various philosophers of science, such as Karl Popper and Thomas Kuhn (Kuhn, 2002), every form of

knowledge is built within cultural and historical paradigms, and this applies to AI as well.

We must beware of a "technocracy of salvation," in which the machine promises redemption and absolute control. AI can assist us but cannot replace us, just as religion must not absolve us of responsibility; no entity, divine or artificial, can relieve us of our essential human nature, defined by errors, doubts, and choices.

Without personal freedom, understood as awareness, creativity, and responsibility, there can be no true spiritual experience. This is clearly underscored by the term "spirit," which originally means "wind," the freest natural element (in Latin it is "*spiritus*", in Greek "*pneuma*", in Hebrew "*ruah*", and in Sanskrit "*atman*"), thereby marking the very close connection between spirituality and freedom. Without freedom, there is no spirituality.

Addressing the subject of AI correctly and effectively is currently one of the major challenges in the field of communication, both scientific and otherwise. The rapid spread of generative AI in recent years was likely unexpected, particularly for those outside the field, and it affects human beings more than we realize, especially considering the possibility that this artificial reality might escape human control.

Religious institutions are also part of this dialogue, even though they are aware that their relationship with scientific knowledge has long been characterized by an ambivalent reality, poised between dialogue, acceptance, conflict, and collaboration.

The Vatican has engaged in communicating about AI, particularly through the words of Pope Francis and articles in the papal daily newspaper *L'Osservatore Romano* (*L'Osservatore Romano*, 2026). Some have addressed this topic in recent years, particularly since 2018, while also actively engaging in understanding and regulating it, for instance through the "Rome Call for AI Ethics" as early as 2020 (*The Call for AI Ethics*, 2026).

The adopted approach has been as follows: first, through the actual production of articles by *L'Osservatore Romano*; and second, through an increasing variety of content offering in-depth analysis, spanning a wide range of formats, from news updates and detailed event reports to reviews of books on AI and reflective pieces.

Over time, the Vatican's approach to the subject has been progressively enriched with fundamental concepts, key terms, and stated positions, leading to the development of descriptive and explanatory content as well as examples of practical AI applications.

The central theme of the Vatican's discourse is ethical reflection, accompanied by a strong call for the necessary regulation of this technology – specifically, the

concept of "algorethics," which originated within the Holy See.

Pope Leo XIV has recently spoken regarding AI, urging that the good of humanity not be sacrificed for private interests and emphasizing the need for an established ethic to accompany the use of these new technological tools, which are currently concentrated in the hands of a few individuals who are difficult to keep in check.

Digital technologies must remain at the service of life and peace (*Vatican News*, 2026). In the contemporary world, it is impossible to ignore the impact of AI on the way society lives; however, its intrusiveness into every social sphere and into personal privacy requires adequate safeguards (*Magnifica Humanitas,* 2026).

The speed of AI innovation prompts us to ask whether genuine control will be possible in the near future, preventing humans from becoming victims of their own inventions. Decisions concerning human life cannot be entrusted to machines.

The Pope highlights this current "worrying situation" and expresses approval for meetings and debates with IISA experts (Address of His Holiness to Members of IISA, 2026), aimed at "reaffirming the irreplaceable nature of human intelligence in the face of technological tools."

The Pope reiterates the need to safeguard the human person, calling for prudence and rigorous checks; any potential slowing down in the adoption of AI should not be viewed as an action against progress, but rather as a firm commitment to exercising responsible care for the human family.

It is therefore vital to define a shared ethic regarding the use of these tools, which unfortunately are currently concentrated in the hands of major economic and technological players that are difficult to control.

We must act to ensure that the good of the human family is not sacrificed for private interests, and to humanize digital technologies so that they serve life, peace, and the integral development of the human being and the community. In the age when machines are beginning to speak, the value of those who speak to the human soul becomes decisive.

Conclusion

The future of AI is not yet written; it will depend on the choices our society decides to make. The philosophy of mind can play a decisive role in guiding these choices, offering a conceptual and ethical frame of reference in our era

of profound and continuous transformation. At stake is not merely the future of machines but, far more importantly, the future of the human being, of our deep identity, our freedom, and our meaning.

Despite the tangible and rapid pace of technological progress, fundamental questions regarding the nature of the mind, consciousness, and the ultimate meaning of life remain open and, in many respects, unresolved.

Philosophy is a tool that enables us to properly evaluate not only what is possible to do, but also, and above all, what is right and meaningful to do; it can help us to grasp the distinction between simulation and reality, imitation and knowledge, and functionality and meaning. Consequently, interaction among philosophy, theology, and science becomes essential in such contexts.

The profound significance of AI extends beyond scientific and computer science theory, carrying major ethical, cultural, and social responsibility implications, especially given its progressive integration into daily life, spanning areas from work to education, and from medicine to the justice system.

It is essential to properly understand the role of the human being in controlling intelligent systems in order to prevent technological dependency, discrimination, and cognitive manipulation, emphasizing the rights and protections that must be guaranteed to users.

This leads us to consider that AI is not merely a matter of software, hardware, computing, or advanced technology, but a new worldview, capable of redefining, even in decidedly negative ways, what it means to be human, what truly distinguishes us from machines, and how we intend to coexist with them in the present and the near future.

Future research directions include the analysis and potential of hybrid cognitive architectures, the study of human-machine interaction and potential transformations of personal identity, political and ethical-moral reflection on the power of intelligent technologies, and the comparison between philosophical ontologies and computational models of mind, areas where a critical and interdisciplinary approach is essential.

We must accept the complexity of the problem and continue to question, with both logical rigor and creative imagination, the meaning of intelligence, consciousness, and life in a world that seems to become increasingly artificial and far from the ethical and moral values of human life.

Unlike weak AI, which makes no claims to consciousness or understanding, strong AI raises radical questions about the possibility of creating genuine artificial minds endowed with awareness, intentionality, and mental states of their own.

From this perspective, philosophical positions diverge profoundly: on one

hand, there are those who defend the possibility that a machine can truly think (such as Dennett), while on the other hand, there are those who maintain that genuine understanding is the exclusive prerogative of biological consciousness (such as Searle, Dreyfus, and Faggin), alongside positions that straddle the two viewpoints, such as that of Chalmers.

Current evidence demonstrates that the mind cannot be reduced to an algorithm, that consciousness is subjective, that experience is embodied, and that relationality and the phenomenological dimension are essential elements for truly understanding what "intelligence" means. The limitations of current symbolic models, along with the gap between simulation and genuine understanding, suggest that today's AI systems remain mindless machines.

However, even in its weak form, AI today makes decisions, examines information, assists in medical diagnoses, determines financial outcomes, and influences opinions as well as both individual and social behaviors. These are areas that touch upon political, legal, and ethical issues, concerning human rights, transparency, and accountability, as well as injustices and inequalities.

These are questions that will long remain at the heart of philosophical and scientific debate, and which elevate the field of AI to one of the most fertile arenas for reflecting on the present and imagining the future on a global scale.

Maintaining focus and engaging in deep reflection have become difficult in the age of AI; they require mental effort, and AI presents formidable competition. Sustained attention and reflection tire the brain, whereas AI makes it easy to satisfy, or convince oneself that one has satisfied, the associated needs. We must build an educational framework that fosters not only digital skills but also critical thinking, ethical awareness, and the capacity for discernment.

This intellectual lethargy should raise concerns about the soundness of our culture and knowledge, highlighting the risks of an abdication of ethical and moral responsibility, as well as the illusion of an algorithmic truth devoid of contradictions. We must preserve the central role of human beings in ethical discernment, rather than stripping them of their capacity for deliberation.

While religion invokes a metaphysical, transcendent, and mysterious entity, AI is a human artifact, immanent and, at least in theory, fully open to inspection. To attribute an aura of infallibility to it is to ignore its true nature.

Ultimately, the future will depend neither on the power of AI nor on the solidity of dogmas, but on our ability to remain critical, responsible, and human, without attributing a higher ontological value to AI; human consciousness and responsibility must remain supreme.

Therefore, the question is whether or not AI fosters personal freedom; the

answer tends to be negative, because freedom arises from work performed firsthand. Conversely, if something else does the work for us, providing ready-made answers and steering us wherever it chooses, freedom diminishes and may ultimately vanish. Human beings make choices and decisions, often guided by wisdom or the heart, whereas a machine makes a technical choice; it has no emotions and cannot experience personal feelings.

▣ Conflict of Interest

The author declares no competing interests.

References

Address of His Holiness to members of IISA. (2026, August 28). Vatican. https://www.vatican.va/content/leo-xiv/en/speeches/2026/augus /documents/ 20260828-iisa.html

Baars, B. J. (2005). Global workspace theory of consciousness: Toward a cognitive neuroscience of human experience. *Progress in Brain Research*, *150*, 45–53. https://doi.org/10.1016/S0079-6123(05)50004-9

Block, N. (2011). Perceptual consciousness overflows cognitive access. *Trends in Cognitive Sciences*, *15*(12), 567–575. https://doi.org/10.1016/j.tics.2011.11.001

Bory, P., Natale, S., & Katzenbach, C. (2025). Strong and weak AI narratives: An analytical framework. *AI & Society*, *40*(5), 2107–2117. https://doi.org/10.1007/s00146-024-02087-8

Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies*. Oxford University Press.

Chalmers, D. J. (1995). Facing up to the problem of consciousness. *Journal of Consciousness Studies*, *2*(3), 200–219. https://doi.org/10.1093/acprof:oso/9780195311105.003.0001

Chalmers, D. J. (1997). *The conscious mind: In search of a fundamental theory*. Oxford University Press.

Dennett, D. C. (1987). *The intentional stance*. Bradford Books.

Dennett, D. C. (1992). *Consciousness explained*. Little, Brown and Company.

Di Sia, P. (2021). Industry 4.0 revolution: Introduction. In C. M. Hussain & P. Di Sia (Eds.), *Handbook of smart materials, technologies, and devices* (pp. 1–20). Springer. https://doi.org/10.1007/978-3-030-58675-1_88-1

Di Sia, P. (2023). Symmetry and the nanoscale: Advances in analytical modeling in the perspective of holistic unification. *Symmetry*, *15*(8), Article 1611. https://doi.org/10.3390/sym15081611

Di Sia, P. (2024). On the concepts of time, space, vacuum and domain of investigation among contemporary physics, philosophy and theological reflection. *Journal of Philosophical Theological Research*, *26*(1), 47–66. https://doi.org/10.22091/jptr.2024.10113.2974

Di Sia, P. (2025a). Nanotechnologies among science, ethics and transhumanism. In R. Pisano (Ed.), *Nanoscience & nanotechnologies: Critical problems, science in society, historical perspectives* (pp. 367–390). Springer. https://doi.org/10.1007/978-3-031-85122-3_8

Di Sia, P. (2025b). *The "primordial dynamic space": A representation of the essence of everything* [Preprint]. Preprints. https://doi.org/10.20944/preprints202508.2149.v1

Di Sia, P. (2026). A narrative review on the evolution of artificial intelligence: Origin, current challenges, and an analytical look at the future. *E-methodology*. Advance online publication.

Dreyfus, H. L. (1992). *What computers still can't do: A critique of artificial reason*. MIT Press.

Eysenck, M. W., & Eysenck, C. (2022). *AI vs humans*. Routledge.

Faggin, F. (2022). *Irriducibile: La coscienza, la vita, i computer e la nostra natura* [Irreducible: Consciousness, life, computers, and our nature]. Mondadori.

Faggin, F. (2024). *Oltre l'invisibile: Dove scienza e spiritualità si uniscono* [Beyond the invisible: Where science and spirituality meet]. Mondadori.

Foot, P. (2003). *Moral dilemmas: And other topics in moral philosophy*. Oxford University Press.

Kuhn, T. S. (2002). *The road since Structure: Philosophical essays, 1970–1993, with an autobiographical interview*. University of Chicago Press.

L'Osservatore Romano. (n.d.). Retrieved September 13, 2026, from https://www.vatican.va/content/osservatore-romano/it.html

Magnifica humanitas. (2026, May 15). Vatican. https://www.vatican.va/content/leo-xiv/it/encyclicals/documents/20260515-magnifica-humanitas.html

Merleau-Ponty, M. (2013). *Phenomenology of perception* (D. A. Landes, Trans.). Routledge. (Original work published 1945)

Metzinger, T. (2010). *The ego tunnel: The science of the mind and the myth of the self*. Basic Books.

Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, *3*(3), 417–424. https://doi.org/10.1017/S0140525X00005756

Sheikh, H., Prins, C., & Schrijvers, E. (2023). Artificial intelligence: Definition and background. In *Mission AI: Research for policy* (pp. 15–41). Springer. https://doi.org/10.1007/978-3-031-21448-6_2

The call for AI ethics. (n.d.). Rome Call for AI Ethics. Retrieved September 13, 2026, from https://www.romecall.org/the-call/

Tononi, G., Boly, M., Massimini, M., & Koch, C. (2016). Integrated information theory: From consciousness to its physical substrate. *Nature Reviews Neuroscience*, *17*(7), 450–461. https://doi.org/10.1038/nrn.2016.44

Vatican News. (2026, August 28). *Papa: L'intelligenza artificiale sia per il bene dell'umanità, non per interessi privati*. https://www.vaticannews.va/it/papa/news/2026-08/papa-intelligenza-artificiale-bene-umanita-interessi-privati.print.html

Zuboff, S. (2019). *The age of surveillance capitalism: The fight for a human future at the new frontier of power*. PublicAffairs.

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

Ethical Authorship in the Age of Generative AI: A Digital Humanist Framework for Scholarly Accountability

Reza Rezvani [1] **A. Mehdi Riazi** [2]

1. Department of Foreign Languages and Linguistics, Shiraz University, Shiraz, Iran, **(Corresponding author)** rezvanireza@gmail.com

2. Macquarie University, Sydney, Australia, mehdi.riazi@mq.edu.au

Abstract

Original Research

The integration of generative artificial intelligence (GenAI) into scholarly publishing has precipitated a philosophical crisis in authorship, agency, and moral responsibility. While a pragmatic consensus holds that AI cannot be an author (Nature Portfolio, 2023), this prohibition remains philosophically under-justified, and no constructive normative framework for human–AI authorship has been offered. This paper examines how traditional humanism, posthumanist distributed agency, and reactive policy pragmatism fail to address the ethical and epistemic challenges of human–AI collaboration. Through philosophical analysis, it investigates whether scholarly authorship can be reconceived to preserve human moral agency and epistemic trust. Engaging with philosophy of technology debates on mediation and moral agency, we argue that digital humanism offers a necessary normative recalibration. This framework reconceptualizes authorship not as the sole origin of text, but as a non-delegable ethical practice of custodial responsibility, where human scholars retain ultimate accountability for AI-augmented work. The paper's principal result is twofold: a philosophical grounding for the authorship prohibition, and custodial authorship as an operational ethical model. We propose that scholarly publishing adopt proactive, value-driven policies enforcing transparency and human oversight.

Keywords:

generative AI, digital humanism, authorship, moral agency, philosophy of technology.

Received: 2025/10/16; **Received in revised form:** 2026/03/14; **Accepted:** 2026/04/12; **Published online:** 2026/04/18

Rezvani, R. & Riazi, A. M. (2026). Ethical Authorship in the Age of Generative AI: A Digital Humanist Framework for Scholarly Accountability. *Journal of Philosophical Theological Research*, (Special Issue on AI, Consciousness, and Human Agency), 28(3), 125-154. https://doi.org/10.22091/jptr.2026.16977.3635

Copyright © The Authors

Introduction

Research integrity has long served as the fundamental pillar ensuring the credibility, trustworthiness, and ethical standards of research and scholarly publishing across disciplines. It underpins the scholarly endeavor by safeguarding the authenticity of the research process and research outputs, and by maintaining public confidence in academic work. However, the rapid evolution of artificial intelligence (AI) technologies – especially GenAI tools like ChatGPT – poses profound challenges to conventional notions of research integrity, including authorship, responsibility, and ethical attribution (Acosta-Enriquez et al., 2024; Hosseini et al., 2024). As AI tools increasingly contribute to creating, editing, and disseminating research and scholarly outputs, the very concept of authorship and what constitutes genuine scholarly contribution is being fundamentally reshaped. This transformation aligns with the fourth digital revolution (Floridi, 2014), which simultaneously reconfigures our knowledge frameworks (epistemology) and modes of existence (ontology) by altering how we perceive and construct reality (Pezzano, 2024). This is not merely a practical challenge, but a deeply philosophical and ethical one, forcing us to rethink what it means to be an author and to be morally responsible for knowledge in an age of machine collaboration.

We are facing an ethical dilemma about how we should deal with AI tools in research and scholarly dissemination. Following ChatGPT's public release in November 2022, researchers rapidly adopted GenAI tools for different purposes. Surveys as early as 2023 indicate approximately 15% of researchers now use such tools to craft grant proposals, while nearly one-third employ them for manuscript drafting assistance (Van Noorden & Perkel, 2023). This trend has intensified, as indicated by Khalifa and Albadawy (2024). They identify main areas where AI supports academic writing and research by facilitating idea generation and research design, enhancing content and structure, streamlining literature review and synthesis, improving data management and analysis, assisting with editing and publishing tasks, and aiding communication and ethical compliance efforts. The scope of adoption became even clearer in a 2025 survey by Naddaf (2025), which examined researchers' reliance on tools like ChatGPT and DeepSeek. Findings underscored a growing consensus that most researchers now view AI as indispensable to scientific research and publishing, with over 50% reporting its superiority to humans in tasks like literature summarization, plagiarism detection, and citation management. Notably, half of the respondents predicted widespread AI adoption within two years, signaling a near-universal shift in scholarly practices.

This swift use of AI tools has ignited debate over GenAI's legitimacy, particularly when researchers list tools like ChatGPT as co-authors (Pavlik,

2023; Prillaman, 2024; Stokel-Walker, 2023). On the other hand, the time-saving potential is substantial, especially in peer review where an estimated 15 million hours are spent annually evaluating rejected papers (American Journal Experts, 2018). Publishers and journals have automated some publishing tasks – including author-reviewer correspondence, plagiarism detection, statistical checks, methodological validation, reviewer identification, and initial quality control – to enhance efficiency (Checco et al., 2021; Kousha & Thelwall, 2023; Lin et al., 2021; Price & Flach, 2017). These developments accelerate predictions that AI could identify "inconsistencies difficult for reviewers to spot" by 2030 (Burley & Moylan, 2017, p. 3).

While studies acknowledge GenAI's benefits for academic tasks (see, e.g., Kasneci et al., 2023; Kohnke et al., 2023; Shen et al., 2023), others document its value for research and writing activities (see, e.g., Dergaa et al., 2023; Hutson, 2022; Lendvai, 2025; Moorhouse & Nejadghanbar, 2025; Morocco-Clarke et al., 2024; Nguyen-Trung et al., 2024; Thorp, 2023; van Dis et al., 2023). Yet these tools also raise pressing ethical questions about permissible uses. As AI applications grow more sophisticated, fundamental questions about authorship and responsibility intensify and demand careful attention. These questions are, at their core, philosophical, touching on moral agency, ethical accountability, and the nature of scholarly contribution. Although the problem is widely acknowledged, its philosophical dimensions remain underexamined. Current scholarship documents that AI fails authorship criteria, but rarely explains why those criteria carry moral weight, nor does it offer a constructive normative framework for human–AI authorship. This paper addresses that gap.

Increasing reliance on AI may jeopardize core principles of genuine authorship, critical thinking, original insight, and moral accountability, since AI-generated texts might lack these inherently human qualities. This ambiguity complicates intellectual attribution and raises legal and ethical concerns about AI-produced content. Consequently, as Casal and Kessler (2023) observed, the scholarly community – including publishers, independent journals, editors, reviewers, and researchers alike – must confront pivotal issues regarding AI's legitimate role in authorship, transparency requirements, and ethical boundaries. In light of this, Berdejo-Espinola and Amano (2023), Hosseini et al. (2024), and Lund and Naheem (2024) have highlighted intense debate over whether AI systems like ChatGPT should be recognized as authors, underscoring the need for clearer standards for ethical responsibility, transparency, and governance. These discussions frame the urgency of adopting governance structures grounded in ethical principles as AI becomes more embedded in scholarly workflows.

Positioned within the philosophy of technology, the primary aim of this

paper is to explore the ethical implications of AI's increasing role in scholarly publishing, particularly how it challenges conventional notions of authorship, responsibility, and integrity. We ask: How can scholarly authorship be philosophically reconceived to preserve human moral agency and epistemic trust in an era of AI collaboration? To address this, we engage in philosophical analysis, examining how different theoretical frameworks – namely humanism, posthumanism, and digital humanism – address the crisis of moral agency and accountability in technologically mediated knowledge production. We seek to clarify the principles of these perspectives, emphasizing digital humanism as a framework that balances innovation with ethical accountability in an era of technological augmentation. Our goal is to provoke dialogue on the adequacy of current standards and detection tools for protecting authentic scholarly effort, and to propose balanced policies that acknowledge AI's role while upholding core ethical values and moral responsibility. Ultimately, we offer a normative philosophical framework for re-establishing authorship on firmer ethical ground. The paper begins by examining debates around authorship in the AI age, then explores policies governing the use of AI tools in scholarly publications, followed by a philosophical examination of the theoretical perspectives underpinning these concepts. The paper concludes with reflections on scholarly publishing policies regarding authorship, primarily informed by digital humanism, to promote ethically responsible practices in this evolving field.

Authorship in the Age of GenAI: Current Debates and Challenges

The transformative impact of AI technologies, particularly large language models (LLMs) like ChatGPT, has ushered in a new era in scholarly writing for publication. Since ChatGPT's public release, there has been swift and widespread adoption across academic and research communities, prompting efforts to regulate and mitigate AI's influence on scholarly work. According to Gulumbe (2024), many publishers and institutions are actively exploring measures to address the ethical and practical challenges posed by AI-generated content. Some researchers (e.g., Islam & Greenwood, 2024) advocate for the outright prohibition of AI use to preserve research integrity. In line with this, and according to Thorp (2023), journals like *Science* have explicitly stated that AI models will not be credited as authors in their publications, emphasizing that any AI-generated content must be properly documented. Failing to cite such contributions may constitute plagiarism (Rezvani, 2024) according to some journals. Similarly, Nature (Nature Portfolio, 2023) has stated that:

> Large Language Models (LLMs), such as ChatGPT, do not currently satisfy our authorship criteria. Notably, an attribution of authorship carries with it accountability for the work, which cannot be effectively applied to LLMs. The use of an LLM should be properly documented in the Methods section (or, if a Methods section is not available, in a suitable alternative part) of the manuscript.

Another perspective (see, e.g., Gulumbe et al., 2024) promotes the constructive integration of GenAI, arguing that restricting these tools is impractical given their pervasive role in modern academia. A more radical perspective, exemplified by Masters (2025), argues that academic authorship (more specifically, the act of writing and composing) itself is a mechanical blending of cultural artifacts – a task ideally suited to LLMs. Drawing on Barthes' (2010) essay, "The Death of the Author," Masters contends that researchers can relinquish authorship to AI entirely, focusing solely on research while AI handles the mechanical act of writing. This view positions authorship as a dispensable proxy for research contribution. Its plausibility, however, depends on an extension of Barthes' argument that Barthes himself did not make. Barthes' claim concerns the interpretive authority of the reader: that a text's meaning is not exhausted by authorial intention. Whether that claim can be extended to the moral responsibility of the writer is a separate question, and one that Masters leaves largely unaddressed. The distinction matters because textual generation and authorial responsibility are not the same thing. Producing sentences can be considered a mechanical act, whereas standing behind a knowledge claim is an ethical one. If authorship is reduced to textual production, it becomes difficult to say who is answerable for the truth, originality, or consequences of what is published; it is precisely this accountability that scholarly integrity presupposes.

As Hammad (2023) observed, while current LLMs excel at mimicking human-like text and organizing content, their capacity for originality and creative thought remains disputed. Whether AI can generate genuinely novel ideas, or merely recombine existing patterns in novel-seeming ways, is an open philosophical question. What is less contentious is that AI cannot assume responsibility for the novelty it produces: even if an AI-generated insight proves original, no machine can be held accountable for its truth, significance, or consequences. In line with Hammad's observation, Salvagno et al. (2023) argue that using GenAI tools like ChatGPT to draft scientific papers still requires expert oversight, as generated texts often lack the nuanced depth and understanding characteristic of human-authored work. Moreover, AI's ability to subtly rephrase or alter content presents unprecedented challenges for plagiarism detection. Discerning these manipulations and maintaining textual

authenticity often requires human judgment, which is increasingly difficult given the sophistication of AI outputs. Studies show that humans are limited in this effort. Gao et al. (2023) found that reviewers could identify only 70% of ChatGPT-generated abstracts, while Weber-Wulff et al. (2023) document significant inaccuracies across detection methods. Adding to this complex situation, Carobene et al. (2024) contend that increasing workloads and shrinking time resources make detection more time-consuming and less reliable. This situation underscores the urgent need for clear ethical guidelines on AI use (Liebrenz et al., 2023).

As Yeo (2025) argues, the rise of GenAI coincides with a crisis in defining scholarly impact, one where traditional metrics like productivity increasingly overshadow human accountability and ethical responsibility (Mazzoleni & Ambrosino, 2024; Moorhouse & Nejadghanbar, 2025). This tension mirrors the core ethical challenge to research integrity posed by GenAI, namely, the reconfiguration of authorship. Specifically, whether non-human entities can hold authorial rights or bear the moral and scholarly obligations inherent in scholarly content creation remains contentious (Mazzi, 2024; Perkins & Roe, 2024). This debate was foreshadowed by historical precedents like the 1981 publication crediting the computer program Racter as an author and the 2019 algorithm-generated book *Lithium-Ion Batteries* by "Beta Writer" (Lee, 2023).

Academic authorship criteria fundamentally focus on accountability and ethical responsibility for intellectual content and integrity. Consequently, the prevailing consensus is clear: AI tools like chatbots do not meet authorship standards because they cannot assume moral or legal responsibility for their output. As Lee (2023, p. 2) notes, a remaining question is: "What about more advanced chatbots in the future?" Considering this question, major publishers explicitly exclude LLMs from authorship eligibility, mandating transparency about AI use in methods or acknowledgments sections. Despite broad acceptance, ethical complexities and questions of moral agency persist, driving calls for standardized ethical guidelines on responsible AI use (Perkins & Roe, 2024; Tate et al., 2023). This prevailing consensus, however, remains a pragmatic policy position rather than a settled philosophical answer. The persistent ethical complexities and fundamental questions of moral agency underscore the need for the conceptual clarity that a philosophical framework like digital humanism aims to provide.

Policies on the Use of AI Tools

Existing research on GenAI policies reveals disciplinary inconsistencies.

Analyses show only 24% of publishers and 87% of top journals have public GenAI policies (Ganjavi et al., 2024), and more recent empirical audits confirm this fragmentation. As a case in point, Bhavsar et al. (2025) found that only 34.6% of STM publishers had publicly available AI policies as of late 2023, while Goyanes et al. (2025) documented that 52.3% of top social science journals do not mention AI use at all in their author guidelines.

These policies commonly prohibit AI authorship and mandate disclosure, but lack consistency in implementation (e.g., where and how to disclose). Science, Technology, Engineering, and Mathematics (STEM) fields, such as cardiology, show greater coherence, with uniform prohibitions against AI authorship, AI-generated images, and AI in peer review (Inam et al., 2024). Perkins and Roe's (2024) cross-disciplinary analysis confirms consensus on permitting documented GenAI use while taking account of its limitations and biases, alongside an emphasis on human ethical accountability. Yet Goyanes et al. (2025) show significant disciplinary variation: communication and sociology demonstrate greater openness to AI use, while economics remains more conservative – a disparity that underscores the absence of cross-disciplinary consensus.

Broadly, three stances underpin these policies: (1) the strict prohibition of AI authorship due to its lack of consciousness, intentionality, and moral agency (see, e.g., COPE Council, 2019; Hosseini et al., 2024); (2) the principle that human authors are ultimately ethically responsible for content, as GenAI operates via pattern recognition, not original intellectual decision-making (Tang et al., 2024); and (3) mandatory disclosure regarding AI use, though the specifics of disclosure remain contested (Perkins & Roe, 2024). These stances affirm that authorship requires capacities like liability for ethical and legal misconduct, which, according to Hosseini et al. (2024), AI cannot fulfill.

The policy fragmentation indicates that pragmatic rules overrun conceptual coherence, leaving fundamental ethical questions about authorship unexamined. Current approaches correctly pinpoint tensions but treat authorship and moral accountability as self-evident, amplifying practical challenges like inconsistent enforcement and detection difficulties (Gulumbe et al., 2024). Consequently, prevailing prohibitions against AI authorship often remain reactive and rooted in practicality, rather than in ethical principle. This gap between high-level ethical principles and implementable practices reflects a broader challenge in AI governance, where translating ethical guidelines into actionable ethical frameworks remains difficult (Papagiannidis et al., 2025).

These persistent tensions between AI's utility and its (in)ability to satisfy authorship criteria demand more than policy adjustments; they require rigorous interrogation of the philosophical and ethical foundations of scholarly

authorship and accountability. In what follows, we briefly explore three pivotal perspectives – namely, humanism, posthumanism, and digital humanism – to establish the groundwork for reimagining ethical authorship in the age of AI.

Conceiving Authorship from Humanism, Posthumanism, and the Digital Humanist Alternative

1. Humanism: Conceptual Diversity and Foundational Principles

The concept of humanism is inherently fluid, adapting across different historical contexts and cultural settings (Prem, 2024). Scholars like Davies (2008) also emphasize the fundamental diversity of humanism, noting the absence of a single, unique definition. According to Prem (2024), this diversity stems from its broad spectrum of historical origins, variously attributed to antiquity, the Italian Renaissance, or German pedagogical traditions. Consequently, contemporary discussions often indiscriminately apply the humanistic label to any philosophy centered on human values, generating substantial conceptual ambiguity (Fuchs, 2022; Kristeller, 1979). Amid this diversity in perspective, the International Humanist and Ethical Union (IHEU, n.d.; see also Copson, 2015, p. 6) provides a unifying modern definition:

> Humanism is a democratic and ethical life stance, which affirms that human beings have the right and responsibility to give meaning and shape to their own lives.

Despite the diversity of interpretations of humanism, certain foundational principles have consistently underpinned its evolution. Its core values focus on acknowledging the dignity of human beings and their ethical responsibility for meaning-making and life-shaping, emphasizing rational inquiry and the pursuit of human flourishing (Copson, 2015). Humanism has also been characterized by a particular interest in understanding human nature through essentialist frameworks that seek to identify qualities unique to humans – distinguishing them from animals and mechanistic views – while also advocating for self-determination and personal autonomy (Prem, 2024).

Ultimately, humanism constitutes "a set of presuppositions that assigns to human beings a special position in the scheme of things," which is "not just a school of thought or a collection of specific beliefs," but "rather a general perspective from which the world is viewed" (Audi, 2015, p. 475). In light of humanist thought, informed by these foundational principles, we can understand human agency as a central principle. Within this paradigm, the

production of scholarly knowledge is understood as fundamentally a human endeavor, emphasizing the intrinsic agency of individuals and their accompanying moral responsibilities, rather than attributing this agency to non-human entities such as machines or emerging generative AI technologies. This perspective aligns with the context in which humanism was initially developed, framing the generation of meaning and knowledge as a uniquely human enterprise endowed with both rights and ethical duties. Consequently, research and scholarly publication are seen as inherently human activities. Thus, individuals possess the right and responsibility to create meaning, including scholarly productions, and to utilize the results of these scholarly endeavors to shape their own lives and experiences (Copson, 2015). The implications for authorship follow directly from this account: if human dignity consists in the capacity for reasoned judgment and moral choice (Copson, 2015), then authorship is not merely a convention of academic publishing, but an expression of that very capacity. To author is to exercise reason, judgment, and accountability – the constitutive features of human dignity as humanism conceives it. Authorship therefore carries moral weight that cannot be transferred to a machine, because machines cannot exercise the capacities that make authorship a moral practice.

2. Posthumanism: Reimagining the Human in a Technological Age

In direct contrast to the humanist position outlined above, posthumanism emerges from a profound critique of humanism, which, according to Ranisch and Sorgner (2014), has "lost its credibility" (p. 16); as Badmington (2000, p. 9) stated, a "crisis in humanism is happening everywhere." This critique has generated diverse posthumanist approaches that, while varying in focus, collectively envision a future where humanism persists only as a residual trace (Chatelier, 2017; Ranisch & Sorgner, 2014). This philosophical movement rejects the anthropocentric premise that "man is the measure of all things" (Ranisch & Sorgner, 2014, p. 16) and challenges the dualist frameworks deeply rooted in Western tradition – such as human/machine, nature/culture, and organic/inorganic – contending that definitions of human beings are historically contingent and ideologically constructed rather than reflections of a fixed essence (Deretić & Sorgner, 2016; Ranisch & Sorgner, 2014).

Expanding on this perspective, posthumanism often revolves around emerging technologies, prompting a reconsideration of the concept of the "human being" through active engagement and interaction with technological innovation (Ranisch & Sorgner, 2014). For example, as a preliminary way to understand this shift, transhumanism advocates for the radical transformation

of human biological and social capacities through technology – an idea that involves human enhancement or augmentation so profound that it could produce life forms with markedly different characteristics, potentially perceived as fundamentally other than human. This technologically driven evolution is commonly called the posthuman (Ranisch & Sorgner, 2014).

Conceptually, posthumanism reframes technology not as an external tool, but as an intrinsic dimension of human evolution. As Valera (2014, p. 486) argues, it embraces a "symbiotic relationship between people and technology," wherein "as much as people create and determine technology, technology creates and determines people." This symbiosis facilitates "hybridization," which is deemed a "progressive elimination and fluidization of differences" that transcends biological constraints and repositions humans as "embodied in an extended technological world" rather than distinct from it (Valera, 2014, pp. 483–484).

Posthumanism does not herald humanity's literal end, but signals "the end of a particular image of us" (Hassan, 1977, p. 845). Hayles (1999) similarly clarifies that "posthumanity is already here," signifying not human extinction, but "the end of a particular conception of the human" (p. 286). However, some strands envisage humanity's replacement by cyborgs and data-centric systems as an inevitable outcome of scientific progress (Harari, 2017). Distinct from transhumanism, a techno-optimist discourse rejects instrumentalist techno-determinism (Popenici, 2023; Ranisch & Sorgner, 2014). It prioritizes reconceptualization over transcendence, emphasizing ethical reorientation toward non-human agents (animals, machines, ecosystems) and emancipatory politics that challenge anthropocentrism (Soper, 2012; Valera, 2014). The techno-optimist perspective seeks to enhance posthumanity through biological or cybernetic augmentation.

According to Tang et al. (2024), posthumanist thought, informed by new materialist frameworks (see, e.g., Barad, 2007), fundamentally reconfigures agency as an emergent property of human–technological networks (Tang & Cooper, 2024). Within this paradigm, scientific knowledge creation – substantially aided by unprecedented methodological innovation and insight generation – is conceived as a collaborative partnership between humans and GenAI, emphasizing shared agency rather than attributing primacy to either entity (Tang et al., 2024). This collaborative dynamic catalyzes novel forms of partnership, integrating human cognitive capabilities with GenAI's computational power to generate, process, analyze, and synthesize complex information (Tang et al., 2024). Consequently, research outputs will increasingly reflect co-creation between human scholars and AI systems, each contributing specialized functions within the knowledge production process (Tang et al., 2024). However, this distributed model of agency raises critical

questions about where ethical accountability ultimately resides – questions that Kotsis (2025) argues are inadequately addressed by current frameworks, as AI cannot bear epistemic or legal responsibility for content errors, fabricated citations, or methodological flaws. This fundamental deficit in accounting for ultimate responsibility is precisely what the digital humanist framework, upholding human sovereignty, is designed to address.

Despite its emancipatory aims, posthumanism faces ethical critiques. Popenici (2023) warns that uncritical narratives of technological integration can fuel dehumanizing constructs disguised as posthuman or antihuman creeds. Similarly, Soper (2012) cautions against soulless techno-determinism. These concerns resonate with the emergent vision of digital humanism, which explicitly seeks to counterbalance technological determinism (Fuchs, 2022), echoing the idea that "life and its fate must be in our hands, in the hands of humans" (Werthner et al., 2024, p. v) to shape technology, not be shaped by it (Fuchs, 2022). This ethical imperative underscores the need for human moral oversight even within technologically mediated partnerships. While posthumanism usefully deconstructs human exceptionalism, its model of hybrid agency faces a critical philosophical limitation: it cannot adequately resolve the problem of *asymmetric moral accountability* within human–AI networks.

3. Digital Humanism: Centering Human Values in the Digital Transformation

As a proposed resolution to the tensions between humanist tradition and posthumanist innovation, the concept of digital humanism – not to be confused with digital humanities – is gaining increased academic attention. Digital humanism has evolved into an influential academic movement with aims that span political and societal spheres (Prem, 2024). It emerges as a proactive philosophical and ethical framework that seeks to "describe, analyze, and, most importantly, influence the complex interplay of technology and humankind, for a better society and life, fully respecting universal human rights" (Werthner, 2024, p. 123). It aims to respond to serious concerns about the societal impact of digitization, ensuring that technology serves humanity while fully respecting universal human rights (Fuchs, 2022; Prem, 2024; Serrano et al., 2024).. This movement explicitly counters techno-economic determinism by advocating for human-centered technological development and democratic participation in tech governance (Coeckelbergh, 2024; Werthner, 2024). At its core, digital humanism represents "an attempt to place human values and interests at the center of the digital transformation" by addressing ethical, social, and political challenges posed by rapid technological advancement (Serrano et al., 2024, p. 1).

The movement is deeply rooted in Enlightenment and humanist traditions, which emphasize reason, freedom, and equality (Nida-Rümelin & Winter, 2024). These ideals, shaped by the Enlightenment, are foundational to the recognition of universal human rights and the belief in the inherent dignity of all individuals, regardless of background (Nida-Rümelin & Winter, 2024). Digital humanism seeks to revive these principles in the context of modern technology, advocating for digital tools that can enhance human autonomy and ethical responsibility rather than diminish them (Nida-Rümelin & Winter, 2024).

The movement crystallized in 2019 following an international conference at the Vienna University of Technology, which produced the Vienna Manifesto on Digital Humanism as a "call to deliberate and act on technological development" (Prem, 2024, p. 2). This manifesto formalized growing unease about the trajectory of digitization and its dehumanizing effects, particularly through AI-driven automation that threatens human agency (Coeckelbergh, 2024). Drawing inspiration from humanist ideals, as Pezzano (2024) articulated, digital humanism bridges "humanistic ideals rooted in Renaissance Humanism with critical reflections on technological progress" (p. 1). Pezzano continues to argue that by blending these ideals with critical engagement with technology, digital humanism seeks to foster a society where humans remain central to technological progress. The manifesto explicitly links the quest for digital humanism to the intellectual traditions of humanism and the Enlightenment (Werthner et al., 2019). According to Nida-Rümelin and Winter (2024), this approach aims to promote values such as rationality, autonomy, and moral responsibility, serving as a counterbalance to the unchecked proliferation of AI and digitalization.

Central to digital humanism is the principle of human ethical sovereignty over technology, insisting that humans must "keep control over digital technology" (Coeckelbergh, 2024, p. 2) and advocating specifically for "computer-supported decisions rather than entirely machine-made decisions" when people are affected (Prem, 2024, p. 3). This demand for human moral oversight directly challenges autonomous AI systems operating without the meaningful and transparent intervention of humans (for further elaboration on this, see Koeszegi, 2024; Weippl & Schrittwieser, 2024; Werthner et al., 2024). Furthermore, digital humanism calls for embedding humanistic and ethical values directly into technology design to respect inherent human dignity and worth, emphasizing that humans are deserving of agency and autonomy (Winter, 2024). This includes a core principle articulated in the Vienna Manifesto, which explicitly proclaims that "Digital technologies should be designed to promote democracy and inclusion. This will require special efforts to overcome current inequalities and to use the emancipatory potential of digital

technologies to make our societies more inclusive" (Werthner et al., 2019, p. 2).

Prem (2024) observes that while the Vienna Manifesto positions digital humanism within Enlightenment and humanistic traditions, these foundational ideals remain insufficiently elaborated and detailed. As discussed in this article, there should be dialogue and discussion on how to mediate between philosophical traditions – in this case, digital humanism – and contemporary technological realities in practice, particularly in their ethical application to contemporary challenges such as AI governance and scholarly publication.

Despite broad acceptance of its motivations, digital humanism faces notable conceptual tensions. As Prem (2024) states, some dismiss it as an oxymoron that paradoxically combines "human" with "non-human" technological systems, while others criticize its reliance on traditional humanist ideals given their historical associations with colonialism, anthropocentrism, and exclusionary worldviews. However, as an evolving field, it continues to navigate these tensions between enduring humanist traditions and the realities of rapidly advancing technology. Nonetheless, its core aim remains clear and vital: to ensure that the primary purpose of technology – especially emerging digital technologies – is and must be to serve humanity, anchoring innovation in human dignity and ethical responsibility rather than merely technical possibility (Pezzano, 2024).

As suggested in the preceding discussion, digital humanism is positioned as a response to the problem of ultimate responsibility. This claim, however, requires specification. Digital humanism does not deny that AI systems participate in knowledge production; it denies that participation entails responsibility. Agency in the descriptive sense – the capacity to produce effects – is distinct from agency in the normative sense – the capacity to be held answerable. AI may possess the former, but cannot possess the latter, because accountability requires the ability to recognize and respond to reasons (Nida-Rümelin & Staudacher, 2024), to have intended an outcome, and to be subject to normative appraisal (Nida-Rümelin & Staudacher, 2024).

Therefore, for the specific crisis of scholarly authorship, digital humanism provides the necessary normative foundation. It re-grounds the authority and accountability of the author not in solitary genius, but in the human scholar's sovereign role as the ethical custodian of AI-augmented knowledge production. By custodial responsibility, we mean the non-delegable duty of a human agent to verify the accuracy and integrity of AI-assisted content, to authorize its submission as a scholarly claim, and to bear moral and professional liability for its consequences. Crucially, this responsibility is assumed at a determinate moment: the act of submission. Prior to submission, AI use is a private writing practice; at submission, it becomes a public act of accountability.

Having thus established this framework, the next section offers our insights and reflections on its implications for scholarly publishing. While focused on research articles, these reflections are also applicable broadly to scholarly dissemination given widespread AI use. They aim to support ethically responsible practices and are born from our critical engagement with the ideas discussed. These reflections are not exhaustive or prescriptive, but are intended to foster dialogue and provoke thoughtful consideration among all stakeholders regarding the ethical challenges and opportunities GenAI introduces.

Reflections and Recommendations

Throughout history, societies have reacted to disruptive innovations with fear, suspicion, and calls for regulation. Ferraris (2024) traces this tendency to Plato's era, when the introduction of writing sparked concerns about the decline of "live" education – fears ultimately addressed through appeals to ethics. Yet, as Ferraris reminds us, "what incites fear and prompts the call for ethics as a remedy against the unknown is not so much the technical apparatus itself but its novelty" (Ferraris, 2024, p. 4). Applied to AI, this suggests that the task is not to reject the technology, but to shape its development through ethical foresight. As Lanier (2011) observes, approaching emerging technologies with anticipation of their negative consequences allows us to shape a positive future – a point reinforced by Grosswiler (2012, p. 93), who notes that "if you anticipate that a new medium will have negative consequences, the consequences are more likely to be positive."

This perspective aligns with digital humanism, which Nida-Rümelin and Staudacher (2024) describe as an ethics for the digital age that interprets and shapes digital transformation in accordance with humanist philosophy. Its central aim is to guide technological development in a way that remains true to humanist values, and its core concept is human authorship – a notion that has shaped moral and legal systems for centuries by highlighting the "triad of reason, freedom, and responsibility" (Nida-Rümelin & Staudacher, 2024, p. 18). Moral agency, on this account, is grounded in the ability to respond thoughtfully to reasons, which forms the normative basis for ethical behavior.

According to Nida-Rümelin and Staudacher (2024), from a conceptual standpoint, digital humanism involves rejecting both the mechanistic view that humans are essentially like machines, and the animistic view that machines can be considered as (or like) humans. Practically, it asserts that ethical responsibility should never be attributed to AI systems, nor should they be permitted to make autonomous ethical decisions. Instead, responsibility

remains a human moral prerogative; that is, those involved in developing and deploying AI must ensure that moral judgment and accountability stay firmly within the realm of human agency. In light of these principles, it becomes clear that while AI is an incredibly powerful and supportive tool, it should not be entrusted with moral responsibility or ethical decision-making. Rather, AI should be viewed as a tool that supports and enhances human tasks without supplanting human reasoning, ethical judgment, and moral agency. This approach ensures that technological progress aligns with the fundamental virtues of reason, freedom, and responsibility – values that lie at the heart of digital humanism and are essential for the ethical development and use of AI.

Having established this digital humanist orientation, the following sections apply its core principles to the specific ethical challenges of authorship and scholarly publishing. We first consider its implications for equity, then examine how it reframes the concepts of authorship and moral accountability, and finally derive recommendations for proactive ethical policy.

1. Beyond Dystopia: A Digital Humanist Ethics for the AI Age

The relationship between technology and access is not new. Rana (2024) observes that historical innovations like the typewriter shifted academic writing toward specialized audiences, privileging those with access to new tools while marginalizing others. This precedent underscores the critical need to align emerging technologies with democratic and ethical values – a need that digital humanism explicitly addresses by asserting that "digital technologies should be designed to promote democracy and inclusion" (Werthner et al., 2019, p. 2).

In scholarly communication, generative AI holds significant promise for advancing these principles by reducing linguistic barriers and democratizing access (Tang et al., 2024). It can level the playing field for researchers with limited English proficiency through enhanced language editing and clarity (Tang et al., 2024). In a sense, AI may serve as a "golden tongue" – a motif long sought across cultures to overcome communicative barriers (Archaeology News Online Magazine, 2026) – granting scholars who have been marginalized by the dominance of English in academic publishing a voice in a discourse where language proficiency has historically dictated participation. This structural disadvantage is well-documented: non-native speakers spend significantly more time writing papers and receive more than double the rejections of their native-speaking counterparts (Amano et al., 2023), and writing in English is perceived as more difficult and anxiety-inducing (Hanauer & Englander, 2011). This linguistic burden, rooted in colonial and neocolonial power dynamics

(Biesta et al., 2025), extends beyond practical barriers to shape whose knowledge is valued and whose voice is heard in global scholarship (Nature Human Behaviour, 2023). Evidence supports this potential, as a survey by Prillaman (2024) found that 55% of researchers prioritize AI-assisted editing, and 75% of European Research Council grant recipients expect GenAI to "reduce language barriers for non-native speakers" by 2030 (European Research Council, 2023, p. 11). When ethically deployed, such tools can ensure that English writing prowess no longer dictates scholarly reach (Bon et al., 2024; Dwivedi et al., 2023), operationalizing digital humanism's goal of equitable and ethical knowledge dissemination.

However, this democratizing potential is not automatic. As Draude (2024) argues, AI can also automate and amplify existing inequities. Scholars warn that privileged researchers and high-income countries are poised to exploit large language models, potentially widening the research gap (Dergaa et al., 2023; van Dis et al., 2023). Empirical evidence bears this out. Liang et al. (2023) found that GPT detectors consistently misclassify non-native English writing as AI-generated, with a false positive rate of 61.22% compared to near-zero for native writers. Giray et al. (2026) explain the mechanism behind this bias, indicating that detection tools operationalize human writing through metrics such as perplexity and burstiness, which encode monolingual, native-speaker norms. Writers whose English reflects careful, formal composition – precisely what L2 academic writing instruction rewards – are structurally disadvantaged, penalized for linguistic discipline rather than flagged for machine generation. What this bias reveals is how technological tools can inadvertently penalize scholars whose first language is not English – ironically, the very scholars who might benefit most from AI-assisted writing. Grieve et al. (2026, p. 2) frame this as a form of epistemic injustice, where "English-language sources are treated as definitive and scholarship that has been traditionally viewed as non-Western is framed as supplementary or contextual." They argue that the dominance of English in academic publishing perpetuates colonial power structures, silencing voices and knowledges that do not conform to Western-centric norms.

Addressing these inequities requires deliberate, value-driven action. From a digital humanist standpoint, this necessitates involving diverse stakeholders – including end-users, developers, ethicists, and representatives from marginalized communities – in AI development and governance (Serrano et al., 2024). This must be coupled with technical and policy efforts to build systems that actively promote ethical fairness and avoid discrimination (Knees et al., 2024; Schmid, 2024), as well as conscious efforts to diversify citation practices, engage with non-English literature, and include voices from the Global South (Grieve et al., 2026). The path

to equity, therefore, lies not in restricting AI, but in deliberately shaping its development and deployment through inclusive, ethicallyinformed, and value-driven design – a core tenet of digital humanism in practice.

2. Digital Humanism in Practice: Equity, Custodial Authorship, and the Imperative of Transparency

The dual potential of GenAI to foster inclusion while threatening integrity brings us to the core ethical crisis: the erosion of authorship. The capabilities that make AI a powerful tool for accessibility also obscure the source of intellectual contributions, challenging foundational concepts of originality and moral accountability. This ambiguity directly jeopardizes the trust and ethical integrity upon which scholarly publishing depends.

Addressing this crisis requires more than technical fixes; it demands a conceptual, ethical, and normative framework. Digital humanism, with its emphasis on human responsibility and reason, provides this foundation, asserting that authorship must remain an inherently human practice grounded in moral agency and non-delegable ethical responsibility (Coeckelbergh, 2024; Nida-Rümelin & Staudacher, 2024). This directly counters posthumanist visions of hybrid agency (Tang & Cooper, 2024) or the radical "death of the academic author" (Masters, 2025), which conflate the mechanical generation of text with the ethical act of authorship. From a digital humanist perspective, the author is not the originator of every sentence, but the ethical custodian who assumes ultimate responsibility for the work's integrity, validity, and ethical soundness. This conception also echoes recent philosophical reflections on human dignity in the age of AI. As Pope Leo XIV (2026, p. 4) warns, the drive to "translate everything, including the mystery of the person, into data and performance" risks reducing human agency to mere computation. Recognizing that every man and woman possesses an "inalienable dignity" (Pope Leo XIV, 2026, p. 13), the encyclical affirms that the human person is not a passive data point, but an active moral subject whose agency cannot be delegated to machines. Kotsis (2025) similarly argues for a hierarchical authorship taxonomy distinguishing human, AI-assisted, and AI-generated contributions – a classification that complements the custodial authorship model by making levels of human oversight explicit. Custodial authorship operationalizes this principle by locating responsibility not in the act of writing per se, but in the exercise of critical judgment and moral oversight over AI-generated contributions.

This redefined authorship also demands ethical transparency. In an era where text provenance is easily obscured, transparency is not a minor concession, but an ethical and epistemic necessity. It is the mechanism that

makes human ethical responsibility visible and assessable. Current frameworks are inadequate: AI can reproduce work without attribution (Else, 2023), and detection tools are unreliable (Perkins et al., 2024). Tsigaris and Teixeira da Silva (2026) demonstrate that under empirically plausible conditions, AI detectors produce more false accusations than correct identifications, with false discovery rates exceeding 70% when the prevalence of detectable AI use is low. They conclude that "transparent disclosure of AI use, clearer institutional guidelines… represent better strategies than an arms-race dynamic" between detectors and AI systems (Tsigaris & Teixeira da Silva, 2026, p. 8). Therefore, we must move beyond vague acknowledgments. Disclosure must be granular and functional, specifying how AI was used (e.g., for literature synthesis, drafting, or language polishing) and what level of human ethical oversight was applied (e.g., editing, verification, integration). This creates an ethical audit trail for human judgment, fulfilling the scholarly community's need to evaluate the work and attribute moral responsibility correctly.

Operationalizing this requires clear, shared ethical standards. The current landscape is fragmented, with varying publisher expectations causing confusion and inconsistent enforcement (Bergstrom & Ruediger, 2024). Building a common ethical vocabulary – defining terms like "AI-assisted drafting," "AI-generated content," and "human intellectual lead" – is a prerequisite for consistent ethical policy and meaningful interdisciplinary dialogue (Bergstrom & Ruediger, 2024; Sueur et al., 2024). This collaborative, multidisciplinary effort itself is consistent with digital humanism's ethical and synthetic approach (Nida-Rümelin & Staudacher, 2024; Werthner et al., 2024).

We therefore propose that publishers explicitly endorse the use of GenAI for legitimate tasks such as language editing and refinement, conditional upon rigorous ethical transparency. This endorsement should be accompanied by clear caveats: overreliance on AI risks introducing distortions, inaccuracies, or "hallucinations" (Alkaissi & McFarlane, 2023; Hicks et al., 2024; Ji et al., 2022), while the "halo effect" of polished prose may mask substantive flaws (Kreuz, 2025). All stakeholders – including authors, reviewers, and editors – should therefore be equipped with critical ethical awareness and training to scrutinize AI-assisted work. The goal is a culture of sovereign and ethically informed use, in which AI enhances human capability without displacing human critical and ethical judgment.

From Principle to Policy: A Proactive Framework for Governance

Custodial authorship and transparency require proactive, structural support beyond individual commitment. Guided by digital humanism, policymaking must

shift from a reactive stance to an anticipatory and ethically driven endeavor. This aligns with concepts of anticipatory ethics (Brey, 2012) and proactive ethics (Floridi & Sanders, 2005), which emphasize embedding ethical foresight and moral reasoning into the development and governance of technology. The urgency of proactive governance is driven by the regulatory gaps documented in recent empirical studies. For example, Bhavsar et al. (2025) found that 65.4% of STM publishers still lack any AI policy, while Goyanes et al. (2025) revealed that only 47.7% of top social science journals mention AI use at all in their author guidelines – a situation that Kotsis (2025) warns leaves editors, institutions, and authors without clear ethical direction, thereby exposing the scientific record to risks of error, bias, and weakened accountability.

A proactive ethical approach calls for a coalition of stakeholders – including publishers, academic institutions, funding bodies, and ethics committees – working together to co-create evolving ethical standards (Papagiannidis et al., 2025). The goal, as Schiaffonati (2024) asserts, shifts from merely avoiding harm to actively fostering societal good, ensuring AI tools uphold integrity, transparency, and human rights.

Publishers must develop clear, adaptable ethical policies that mandate detailed disclosure and establish verification protocols, requiring authors to certify their understanding of responsible and ethical AI use. These policies must be iterative and inclusive, forged through multidisciplinary and ethically informed partnerships with researchers, ethicists, and technologists. Furthermore, publishers must transparently disclose their own use of AI in editorial and production processes, practicing the reciprocal ethical integrity they demand of authors. This approach is reflected in recent regulatory developments, such as the EU's Digital Omnibus on AI (Regulation 2026/1744), which establishes AI literacy as a legal obligation for providers and deployers (Article 4) and mandates guidelines to minimize compliance burdens while maintaining human accountability (Recital 37). Such frameworks operationalize the digital humanist commitment to human-centric governance.

A critical complication is the inherent opacity of many AI systems (Dwivedi et al., 2023; Tamburrini, 2024; Weippl & Schrittwieser, 2024). A digital humanist stance holds that this technical "black box" does not absolve scholars of ethical responsibility. On the contrary, it heightens the author's duty of vigilant ethical custodianship. Authors are ethically obliged to exercise due diligence, rigorously verify outputs, and bear full moral and professional responsibility for any inaccuracies or misuse, regardless of the tool's internal unknowability (Werthner et al., 2024).

Proactive policymaking is thus the institutional enactment of the digital humanist commitment to human ethical sovereignty. It is the means by which

the scholarly community collectively ensures that technological progress remains anchored in the human values of dignity, equity, and moral accountability, thereby safeguarding the integrity and ethical trust essential to its mission. This represents a conscious philosophical choice: to govern technology through human values, thereby asserting that the trajectory of scholarly communication is a matter of ethical and political will, not an inevitable technical outcome.

Conclusion

The rapid integration of generative AI into scholarly publishing represents more than a practical challenge; it is a profound conceptual and ethical crisis that threatens the very foundations of authorship, moral accountability, and trust upon which academia depends. As this paper has shown, current policy responses are fragmented and reactive (Ganjavi et al., 2024; Perkins & Roe, 2024), prioritizing pragmatic rules over philosophical and ethical coherence. This inadequacy, combined with unreliable detection tools (Perkins et al., 2024; Weber-Wulff et al., 2023) and the inherent opacity of AI systems (Dwivedi et al., 2023; Werthner, 2024), leaves scholarly integrity and ethical trust vulnerable.

In confronting this crisis, we have argued that neither a retreat to traditional humanist essentialism nor an embrace of posthumanist distributed agency provides an adequate ethical framework. Instead, digital humanism offers the essential conceptual and normative ethical foundation. It resolves the tension by asserting three non-negotiable principles for the digital age: the non-delegability of human moral responsibility, the sovereignty of human ethical judgment over technology, and epistemic and ethical transparency as a condition for trust.

Scholarly authorship can preserve human moral agency in an era of AI collaboration only if it is relocated from textual production to moral accountability. From these principles, we have derived a coherent framework for reconstructing authorship. This framework redefines the author not as the sole generator of text, but as the responsible ethical custodian of scholarly work – the agent who exercises critical and moral oversight, verifies content, and assumes ultimate ethical accountability. It mandates granular ethical transparency to make this custodianship visible, and demands proactive, value-driven, and ethically informed governance to institutionalize these norms across the scholarly ecosystem. This approach directly addresses the dual edges of GenAI, harnessing its potential for equity and inclusion while erecting necessary ethical guardrails to protect intellectual integrity.

The contribution of this framework is threefold: it grounds the authorship prohibition in moral philosophy rather than policy pragmatism; it converts that grounding into an actionable model; and it bridges the gap between ethical principle and scholarly practice by translating core digital humanist values – reason, dignity, and responsibility – into policies for transparency and governance. In doing so, it resists both the erosion of ethical responsibility and the myth of hybrid agency, offering instead a normative account of authorship that centers human moral agency in human–AI collaboration.

The framework developed here carries implications for the principal actors in scholarly publishing. For publishers, it implies the development of clear and evolving standards for disclosure and verification. For authors, it implies a conception of authorship as vigilant custodianship – exercising oversight and accepting responsibility for AI-augmented work. For academic institutions and funders, it implies support for this shift through training, policy, and incentives.

The choice is not between adopting or rejecting AI, but between allowing technology to erode scholarly norms and shaping technology to serve enduring human values. If the digital humanist commitment to reason, freedom, and responsibility is taken seriously, generative AI can be integrated into scholarly practice in a way that expands access and preserves the integrity, accountability, and trust on which the scholarly enterprise depends.

▣ Acknowledgements

The authors would like to thank the anonymous reviewers for their constructive feedback on an earlier version of this manuscript.

▣ AI Use Declaration

The authors used ChatGPT (Version 3.5) for language editing and polishing. The authors reviewed and revised all content and take full responsibility for the final manuscript.

▣ Conflict of Interest

The authors declare no competing interests.

References

Acosta-Enriquez, B. G., Arbulú Ballesteros, M. A., Huamaní Jordan, O., López, R. S., Paredes, P. S., & Cruz, J. L. (2024). Analysis of college students' attitudes toward the use of ChatGPT in their academic activities: Effect of intent to use, verification of information, and responsible use. *BMC Psychology, 12* (1), Article 255. https://doi.org/10.1186/s40359-024-01764-z

Alkaissi, H., & McFarlane, S. I. (2023). Artificial hallucinations in ChatGPT: Implications in scientific writing. *Cureus, 15*(2), Article e35179. https://doi.org/10.7759/cureus.35179

Amano, T., Ramírez-Castañeda, V., Berdejo-Espinola, V., Borokini, I., Chowdhury, S.,... & Zahawi, R. A. (2023). The manifold costs of being a non-native English speaker in science. *PLOS Biology*, *21*(7), Article e3002184. https://doi.org/10.1371/journal.pbio.3002184

American Journal Experts. (2018). *Peer review: How we found 15 million hours of lost time*. http://www.aje.com/en/arc/peer-review-process-15-million-hours-lost-time/

Archaeology News Online Magazine. (2026, July 6). Ancient tombs with golden tongues found at Egypt's Marina El Alamein reveal 2,000-year-old burial secrets. *Archaeology News Online Magazine*. https://archaeologymag.com/2026/07/tombs-with-golden-tongues-at-marina-el-alamein/

Audi, R. (Ed.). (2015). *The Cambridge dictionary of philosophy*. Cambridge University Press. https://doi.org/10.1017/CBO9781139057509

Badmington, N. (Ed.). (2000). *Posthumanism*. Palgrave.

Barad, K. (2007). *Meeting the universe halfway: Quantum physics and the entanglement of matter and meaning*. Duke University Press.

Barthes, R. (2010). The death of the author. In V. B. Leitch (Ed.), *The Norton anthology of theory and criticism* (pp. 1322–1326). W. W. Norton & Company.

Bearman, M., & Ajjawi, R. (2023). Learning to work with the black box: Pedagogy for a world with artificial intelligence. *British Journal of Educational Technology, 54*(5), 1160–1173. https://doi.org/10.1111/bjet.13337

Berdejo-Espinola, V., & Amano, T. (2023). AI tools can improve equity in science. *Science, 379*(6636), 991. https://doi.org/10.1126/science.adg9714

Bergstrom, T., & Ruediger, D. (2024). *A third transformation? Generative AI and scholarly publishing* (Research Report). Ithaka S+R. https://doi.org/10.18665/sr.321519

Bhavsar, D., Duffy, L., Jo, H., & Sokolova, A. (2025). Policies on artificial intelligence chatbots among academic publishers: A cross-sectional audit. *Research Integrity and Peer Review*, *10*, Article 1. https://doi.org/10.1186/s41073-025-00158-y

Biesta, G., Takayama, K., Kettle, M., & Heimans, S. (2025). How to say in English what you cannot say in English? Dilemmas of 'global' scholarship and small steps forward. *Asia-Pacific Journal of Teacher Education*, *53*(2), 135–137. https://doi.org/10.1080/1359866X.2025.2469970

Bon, A., Saa-Dittoh, F., & Akkermans, H. (2024). Bridging the digital divide. In H. Werthner, C. Ghezzi, J. Kramer, J. Nida-Rümelin, B. Nuseibeh, E. Prem, & A. Stanger (Eds.), *Introduction to digital humanism: A textbook* (pp. 283–298). Springer Nature. https://doi.org/10.1007/978-3-031-45304-5_19

Brey, P. A. (2012). Anticipatory ethics for emerging technologies. *NanoEthics, 6*(1), 1–13.

Burley, R., & Moylan, E. (2017). *What might peer review look like in 2030? A report from BioMed Central and Digital Science* (SpotOn Report). BioMed Central & Digital Science. https://doi.org/10.6084/m9.figshare.4884878

Carobene, A., Padoan, A., Cabitza, F., Banfi, G., & Plebani, M. (2024). Rising adoption of artificial intelligence in scientific publishing: Evaluating the role, risks, and ethical implications in paper drafting and review process. *Clinical Chemistry and Laboratory Medicine, 62*(5), 835–843. https://doi.org/10.1515/cclm-2023-1136

Casal, J. E., & Kessler, M. (2023). Can linguists distinguish between ChatGPT/AI and human writing? A study of research ethics and academic publishing. *Research Methods in Applied Linguistics, 2*(3), Article 100068. https://doi.org/10.1016/j.rmal.2023.100068

Chan, C. K. Y. (2023). A comprehensive AI policy education framework for university teaching and learning. *International Journal of Educational Technology in Higher Education, 20*(38), 1–25.

Chatelier, S. (2017). Beyond the humanism/posthumanism debate: The educational implications of Said's critical, humane praxis. *Educational Theory, 67*(6), 657–672. https://doi.org/10.1111/edth.12278

Checco, A., Bracciale, L., Loreti, P., Pinfield, S., & Bianchi, G. (2021). AI-assisted peer review. *Humanities and Social Sciences Communications, 8*(1), 1–11.

Coeckelbergh, M. (2024). What is digital humanism? A conceptual analysis and an argument for a more critical and political digital (post)humanism. *Journal of Responsible Technology, 17*, Article 100073. https://doi.org/10.1016/j.jrt.2023.100073

COPE Council. (2019). *COPE discussion document: Authorship*. Committee on Publication Ethics. https://doi.org/10.24318/cope.2019.3.3

Copson, A. (2015). What is humanism? In A. Copson & A. C. Grayling (Eds.), *The Wiley Blackwell handbook of humanism* (pp. 1–33). Wiley-Blackwell.

Davies, T. (2008). *Humanism*. Routledge.

Deretić, I., & Sorgner, S. L. (2016). Introduction. In I. Deretić & S. L. Sorgner (Eds.), *From humanism to meta-, post- and transhumanism?* (pp. 13–20). Peter Lang.

Dergaa, I., Chamari, K., Zmijewski, P., & Ben Saad, H. (2023). From human writing to artificial intelligence generated text: Examining the prospects and potential threats of ChatGPT in academic writing. *Biology of Sport, 40*, 615–622. https://doi.org/10.5114/biolsport.2023.125623

Draude, C. (2024). Digital transformation through the lens of intersectional gender research: Challenges and needs for action. In H. Werthner, C. Ghezzi, J. Kramer, J. Nida-Rümelin, B. Nuseibeh, E. Prem, & A. Stanger (Eds.), *Introduction to digital*

humanism: A textbook (pp. 83–97). Springer. https://doi.org/10.1007/978-3-031-45304-5_6

Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., Baabdullah, A. M., Koohang, A., Raghavan, V., Ahuja, M., Albanna, H., Albashrawi, M. A., AlBusaidi, A. S., Balakrishnan, J., Barlette, Y., Basu, S., Bose, I., Brooks, L., Buhalis, D., … Wright, R. (2023). "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges, and implications of generative conversational AI for research, practice, and policy. *International Journal of Information Management*, *71*, Article 102642. https://doi.org/10.1016/j.ijinfomgt.2023.102642

Else, H. (2023). Abstracts written by ChatGPT fool scientists. *Nature, 613*(7944), 423. https://doi.org/10.1038/d41586-023-00056-7

European Research Council (2023). Foresight: Use and impact of artificial intelligence in the scientific process. Luxembourg: European Research Council. UNT Libraries Government Documents Department. Retrieved January 10, 2025, from https://digital.library.unt.edu/ark:/67531/metadc2289481/m1/4/

Ferraris, M. (2024). Intelligence as a human life form. *Journal of Responsible Technology, 18*, Article 100081. https://doi.org/10.1016/j.jrt.2024.100081

Floridi, L. (2014). *The fourth revolution: How the infosphere is reshaping human reality*. Oxford University Press.

Floridi, L., & Sanders, J. W. (2005). Internet ethics: The constructionist values of *homo poieticus*. In R. J. Cavalier (Ed.), *The impact of the internet on our moral lives* (pp. 195–214). State University of New York Press.

Fuchs, C. (2022). *Digital humanism: A philosophy for 21st century digital society*. Emerald Publishing.

Ganjavi, C., Eppler, M. B., Pekcan, A., Biedermann, B., Abreu, A., Collins, G. S., Gill, I. S., & Cacciamani, G. E. (2024). Publishers' and journals' instructions to authors on use of generative artificial intelligence in academic and scientific publishing: Bibliometric analysis. *BMJ*, *384*, Article e077192. https://doi.org/10.1136/bmj-2023-077192

Gao, C. A., Howard, F. M., Markov, N. S., Dyer, E. C., Ramesh, S., Luo, Y., & Pearson, A. T. (2023). Comparing scientific abstracts generated by ChatGPT to real abstracts with detectors and blinded human reviewers. *NPJ Digital Medicine*, *6*, Article 75. https://doi.org/10.1038/s41746-023-00819-6

Giray, L., Sevnarayan, K., & Maphoto, K. B. (2026). GPTZero and the challenges of AI detection in assessing writing. *Assessing Writing*, *69*, Article 101078. https://doi.org/10.1016/j.asw.2026.101078

Goyanes, M., Lopezosa, C., & Piñeiro-Naval, V. (2025). The use of artificial intelligence (AI) in research: A review of author guidelines in leading journals across eight social science disciplines. *Scientometrics*, *130*, 3725–3741. https://doi.org/10.1007/s11192-025-05377-0

Grieve, A., de Sales, A., Adarkwah, M. & Rumyantseva, N. (2026). Epistemic (in)justice in higher education publishing. *Journal of University Teaching and Learning Practice,*

23(2). https://doi.org/10.53761/eap

Grosswiler, P. (2012). Editorial. *Explorations in Media Ecology, 11*(2), 91–93.

Gulumbe, B. H. (2024). Obvious artificial intelligence-generated anomalies in published journal articles: A call for enhanced editorial diligence. *Learned Publishing*, *37*(4), 1–5. https://doi.org/10.1002/leap.1626

Gulumbe, B. H., Audu, S. M., & Hashim, A. M. (2024). Balancing AI and academic integrity: What are the positions of academic publishers and universities? *AI & Society, 40*, 1775–1784. https://doi.org/10.1007/s00146-024-01946-8

Hammad, M. (2023). The impact of artificial intelligence (AI) programs on writing scientific research. *Annals of Biomedical Engineering, 51*(3), 459–460.

Hanauer, D. I., & Englander, K. (2011). Quantifying the burden of writing research articles in a second language: Data from Mexican scientists. *Written Communication*, *28*(4), 403–416. https://doi.org/10.1177/0741088311420056

Harari, Y. N. (2017). *Homo Deus: A brief history of tomorrow.* Vintage.

Hassan, I. (1977). Prometheus as performer: Toward a posthumanist culture? *The Georgia Review, 31*(4), 830–850.

Hayles, N. K. (1999). *How we became posthuman: Virtual bodies in cybernetics, literature, and informatics*. University of Chicago Press.

Hicks, M. T., Humphries, J., & Slater, J. (2024). ChatGPT is bullshit. *Ethics and Information Technology, 26*(1), 38. https://doi.org/10.1007/s10676-024-09775-5

Hosseini, M., Rasmussen, L. M., & Resnik, D. B. (2024). Using AI to write scholarly publications. *Accountability in Research, 31*(7), 715–723. https://doi.org/10.1080/08989621.2023.2168535

Hutson, M. (2022). Could AI help you to write your next paper? *Nature*, *611*, 192–193. https://doi.org/10.1038/d41586-022-03479-w

Inam, M., Sheikh, S., Minhas, A. M. K., Vaughan, E. M., Krittanawong, C., Samad, Z., Lavie, C. J., Khoja, A., D'Cruze, M., Slipczuk, L., & Alarakhiya, F. (2024). A review of top cardiology and cardiovascular medicine journal guidelines regarding the use of generative artificial intelligence tools in scientific writing. *Current Problems in Cardiology*, *49*(3), Article 102387.

International Humanist and Ethical Union. (n.d.). *What is humanism?* Retrieved June 2025, from https://humanists.international/what-is-humanism/

Islam, G., & Greenwood, M. (2024). Generative artificial intelligence as hypercommons: Ethics of authorship and ownership. *Journal of Business Ethics, 192*(2), 659–663. https://doi.org/10.1007/s10551-024-05741-9

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2022). Survey of hallucination in natural language generation. *ACM Computing Surveys, 55*(12), 1–38. https://doi.org/10.1145/3571730

Kasneci, E., Sesler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F.,... & Weller, J. (2023). ChatGPT for good? On opportunities and challenges of large

language models for education. *Learning and Individual Differences, 103*, Article 102274. https://doi.org/10.1016/j.lindif.2023.102274

Knees, P., Neidhardt, J., & Nalis, I. (2024). Recommender systems: Techniques, effects, and measures toward pluralism and fairness. In H. Werthner, C. Ghezzi, J. Kramer, J. Nida-Rümelin, B. Nuseibeh, E. Prem, & A. Stanger (Eds.), *Introduction to digital humanism: A textbook* (pp. 417–434). Springer. https://doi.org/10.1007/978-3-031-45304-5_27

Knight, S., Dickson-Deane, C., Heggart, K., Kitto, K., Kozanoglu, D. C., Maher, D., Narayan, B., & Zarrabi, F. (2023). Generative AI in the Australian education system: An open data set of stakeholder recommendations and emerging analysis from a public inquiry. *Australasian Journal of Educational Technology*, 39(5), 101–124. https://doi.org/10.14742/ajet.8922

Koeszegi, S. T. (2024). AI @ work: Human empowerment or disempowerment? In H. Werthner, C. Ghezzi, J. Kramer, J. Nida-Rümelin, B. Nuseibeh, E. Prem, & A. Stanger (Eds.), *Introduction to digital humanism: A textbook* (pp. 175–196). Springer. https://doi.org/10.1007/978-3-031-45304-5_12

Kohnke, L., Moorhouse, B. L., & Zou, D. (2023). ChatGPT for language teaching and learning. *RELC Journal, 54*(2), 1–13. https://doi.org/10.1177/00336882231162868

Kotsis, K. T. (2025). Redefining Scientific Authorship in the Age of AI: Challenges for Editors and Institutions. *European Journal of Innovative Studies and Sustainability, 1*(5), 23-33. https://doi.org/10.59324/ejiss.2025.1(5).03

Kreuz, R. (2025). Too many em dashes? Weird words like "delves"? Spotting text written by ChatGPT is still more art than science. The Conversation. Retrieved February 9, 2025, from https://theconversation.com/too-many-em-dashes-weird-words-like-delves-spotting-text-written-by-chatgpt-is-still-more-art-than-science-259629

Kristeller, P. O. (1979). *Renaissance thought and its sources*. Columbia University Press.

Lanier, J. (2011). *You are not a gadget: A manifesto*. Vintage.

Lee, J. Y. (2023). Can an artificial intelligence chatbot be the author of a scholarly article? *Journal of Educational Evaluation for Health Professions, 20*, 6. https://doi.org/10.3352/jeehp.2023.20.6

Lendvai, G. F. (2025). ChatGPT in academic writing: A scientometric analysis of literature published between 2022 and 2023. *Journal of Empirical Research on Human Research Ethics*, *20*(3), 131–148. https://doi.org/10.1177/15562646251350203

Liang, W., Yuksekgonul, M., Mao, Y., Wu, E., & Zou, J. (2023). GPT detectors are biased against non-native English writers. *Patterns*, *4*(7), Article 100779. https://doi.org/10.1016/j.patter.2023.100779

Liebrenz, M., Schleifer, R., Buadze, A., Bhugra, D., & Smith, A. (2023). Generating scholarly content with ChatGPT: Ethical challenges for medical publishing. *The Lancet Digital Health*, *5*(3), e105–e106. https://doi.org/10.1016/S2589-7500(23)00019-5

Lin, J., Song, J., Zhou, Z., & Shi, X. (2021). Automated scholarly paper review: Possibility and challenges. *arXiv preprint*. https://doi.org/10.48550/arXiv.2111.07533

Lund, B. D., & Naheem, K. T. (2024). Can ChatGPT be an author? A study of artificial intelligence authorship policies in top academic journals. *Learned Publishing, 37*(1), 13–21. https://doi.org/10.1002/leap.1582

Masters, K. (2025). Artificial intelligence and the death of the academic author. *Medical Teacher, 47*(10), 1563–1565. https://doi.org/10.1080/0142159X.2025.2523463

Mazzi, F. (2024). Authorship in artificial intelligence-generated works: Exploring originality in text prompts and artificial intelligence outputs through philosophical foundations of copyright and collage protection. *Journal of World Intellectual Property*, *27*(3), 410–427. https://doi.org/10.1111/jwip.12310

Moorhouse, B. L., & Nejadghanbar, H. (2025). A response to Szudarski's (2025) book review of 'Vocabulary, corpus and language teaching. A machine-generated literature overview'. *ELT Journal*, *79*(3), 490–495. https://doi.org/10.1093/elt/ccaf018

Mazzoleni, S., & Ambrosino, N. (2024). How artificial intelligence is changing scientific publishing? Unrequested advices for young researchers II. *Pulmonology, 30*(5), 413–415. https://doi.org/10.1016/j.pulmoe.2024.04.011

Morocco-Clarke, A., Sodangi, F. A., & Momodu, F. (2024). The implications and effects of ChatGPT on academic scholarship and authorship: A death knell for original academic publications? *Information and Communications Technology Law, 33*(1), 21–41. https://doi.org/10.1080/13600834.2023.2239623

Naddaf, M. (2025, February 4). How are researchers using AI? Survey reveals pros and cons for science. *Nature*. Advance online publication. https://doi.org/10.1038/d41586-025-00343-5

Nature Human Behaviour. (2023). Scientific publishing has a language problem [Editorial]. *Nature Human Behaviour*, *7*, 1019–1020. https://doi.org/10.1038/s41562-023-01679-6

Nature Portfolio. (2023). Artificial intelligence (AI). Retrieved December 20, 2024, from https://www.nature.com/nature-portfolio/editorial-policies/ai

Nguyen-Trung, K., Saeri, A. K., & Kaufman, S. (2024). Applying ChatGPT and AI-powered tools to accelerate evidence reviews. *Human Behavior and Emerging Technologies, 2024*(1), Article 8815424. https://doi.org/10.1155/2024/8815424

Nida-Rümelin, J., & Staudacher, K. (2024). Philosophical foundations of digital humanism. In H. Werthner, C. Ghezzi, J. Kramer, J. Nida-Rümelin, B. Nuseibeh, E. Prem, & A. Stanger (Eds.), *Introduction to digital humanism: A textbook* (pp. 1–20). Springer.

Nida-Rümelin, J., & Winter, D. (2024). Humanism and enlightenment. In H. Werthner, C. Ghezzi, J. Kramer, J. Nida-Rümelin, B. Nuseibeh, E. Prem, & A. Stanger (Eds.), *Introduction to digital humanism: A textbook* (pp. 3–16). Springer.

Papagiannidis, E., Mikalef, P., & Conboy, K. (2025). Responsible artificial intelligence governance: A review and research framework. *Journal of Strategic Information Systems, 34*(2), Article 101885. https://doi.org/10.1016/j.jsis.2024.101885

Pavlik, J. V. (2023). Collaborating with Chatgpt: Considering the Implications of Generative Artificial Intelligence for Journalism and Media Education. *Journalism & Mass Communication Educator, 78,* 84-93. https://doi.org/10.1177/10776958221149577

Perkins, M., & Roe, J. (2024). Academic publisher guidelines on AI usage: A ChatGPT supported thematic analysis. *F1000Research, 12*, Article 1398. https://doi.org/10.12688/f1000research.142411.2

Perkins, M., Roe, J., Postma, D., McGaughran, J., & Hickerson, D. (2024). Detection of GPT-4 generated text in higher education: Combining academic judgement and software to identify generative AI tool misuse. *Journal of Academic Ethics*, *22* (1), 89–113. https://doi.org/10.1007/s10805-023-09492-6

Pezzano, G. (2024). Are we done with (Wordy) manifestos? Towards an introverted digital humanism. *Journal of Responsible Technology*, *17*, Article 100078. https://doi.org/10.1016/j.jrt.2024.100078

Pope Leo XIV. (2026). *Magnifica humanitas: On safeguarding the human person in the time of artificial intelligence* [Encyclical letter]. Vatican City.

Popenici, S. (2023). *Artificial intelligence and learning futures: Critical narratives of technology and imagination in higher education*. Routledge.

Prem, E. (2024). Principles of digital humanism: A critical post-humanist view. *Journal of Responsible Technology*, *17*, Article 100075. https://doi.org/10.1016/j.jrt.2024.100075

Price, S., & Flach, P. A. (2017). Computational support for academic peer review: A perspective from artificial intelligence. *Communications of the ACM*, *60*(3), 70–79. https://doi.org/10.1145/2979672

Prillaman, M. (2024, March 7). Is ChatGPT making scientists hyper-productive? The highs and lows of using AI. *Nature*, *627*(8002), 16–17. https://doi.org/10.1038/d41586-024-00592-w

Rana, N. K. (2024). Generative AI and academic research: A review of the policies from selected HEIs. *Higher Education for the Future, 12*(1), 97–113. https://doi.org/10.1177/23476311241303800

Ranisch, R., & Sorgner, S. L. (2014). Introducing post- and transhumanism. In R. Ranisch & S. L. Sorgner (Eds.), *Post- and transhumanism: An introduction* (pp. 7–28). Peter Lang Edition.

Regulation (EU) 2026/1744 of the European Parliament and of the Council of 8 July 2026 amending Regulations (EU) 2024/1689, (EU) 2018/1139 and (EU) 2023/1230 as regards the simplification of the implementation of harmonised rules on artificial intelligence (Digital Omnibus on AI). (2026). *Official Journal of the European Union*, L 2026/1744. http://data.europa.eu/eli/reg/2026/1744/oj

Rezvani, R. (2024). A qualitative study of TEFL graduate students' source text use: Learning, problems, and strategies. *Teaching English Language, 18*(2), 93-126. https://doi.org/ 10.22132/tel.2024.452802.1592

Salvagno, M., Taccone, F. S., & Gerli, A. G. (2023). Can artificial intelligence help for scientific writing? *Critical Care, 27*(1), 75. https://doi.org/10.1186/s13054-023-04380-2

Schiaffonati, V. (2024). Promises and perils in moralizing technologies. In H. Werthner, C. Ghezzi, J. Kramer, J. Nida-Rümelin, B. Nuseibeh, E. Prem, & A. Stanger (Eds.),

Introduction to digital humanism: A textbook (pp. 255–265). Springer. https://doi.org/10.1007/978-3-031-45304-5_17

Schmid, U. (2024). Trustworthy artificial intelligence: Comprehensible, transparent and correctable. In H. Werthner, C. Ghezzi, J. Kramer, J. Nida-Rümelin, B. Nuseibeh, E. Prem, & A. Stanger (Eds.), *Introduction to digital humanism: A textbook* (pp. 151–164). Springer.

Serrano, G., Striano, F., & Umbrello, S. (2024). Digital humanism as a bottom-up ethics. *Journal of Responsible Technology, 18*, Article 100082.

Shen, Y., Heacock, L., Elias, J., Hentel, K., Reig, B., Shih, G., & Moy, L. (2023). ChatGPT and other large language models are double-edged swords. *Radiology, 307*(2), Article e230163. https://doi.org/10.1148/radiol.230163

Soper, K. (2012). The humanism in posthumanism. *Comparative Critical Studies, 9*(3), 365–378. https://doi.org/10.3366/ccs.2012.0069

Stokel-Walker, C. (2023, January 18). ChatGPT listed as author on research papers: Many scientists disapprove. *Nature, 613(7945), 620–621.* https://doi.org/10.1038/d41586-023-00107-z

Sueur, C., Lombard, J., Capra, O., Beltzung, B., & Pelé, M. (2024). Exploration of the creative processes in animals, robots, and AI: Who holds the authorship? *Humanities and Social Sciences Communications, 11,* Article 611. https://doi.org/10.1057/s41599-024-03125-y

Tamburrini, G. (2024). Artificial intelligence and large-scale threats to humanity. In H. Werthner, C. Ghezzi, J. Kramer, J. Nida-Rümelin, B. Nuseibeh, E. Prem, & A. Stanger (Eds.), *Introduction to digital humanism: A textbook* (pp. 241–254). Springer. https://doi.org/10.1007/978-3-031-45304-5_16

Tang, K. S., Cooper, G., & Nielsen, W. (2024). Philosophical, legal, ethical, and practical considerations in the emerging use of generative AI in academic journals: Guidelines for research in science education (RISE). *Research in Science Education, 54(*5), 797-807. https://doi.org/10.1007/s11165-024-10192-3

Tang, K.-S., & Cooper, G. (2024). The role of materiality in an era of generative artificial intelligence. *Science & Education, 34*, 731–746. https://doi.org/10.1007/s11191-024-00508-0

Tate, T., Doroudi, S., Ritchie, D., Xu, Y., & Warschauer, M. (2023). Educational research and AI-generated writing: Confronting the coming tsunami. *EdArXiv*. https://doi.org/10.35542/osf.io/4mec3

Thorp, H. H. (2023). ChatGPT is fun, but not an author. *Science, 379*(6630), 313. https://doi.org/10.1126/science.adg7879

Tsigaris, P., & Teixeira da Silva, J. A. (2026). AI detecting AI in academic writing: Why most AI detector findings are false. *Research Ethics*, *7*, Article 101396. https://doi.org/10.1016/j.nexres.2026.101396

Valera, L. (2014). Posthumanism: beyond humanism?. *Cuadernos de bioetica : revista oficial de la Asociacion Espanola de Bioetica y Etica Medica*, *25*(85), 481–491.

van Dis, E. A. M., Bollen, J., Zuidema, W., van Rooij, R., & Bockting, C. L. (2023). ChatGPT: Five priorities for research. Nature, 614, 224–226. https://doi.org/10.1038/d41586-023-00288-7.

Van Noorden, R., & Perkel, J. M. (2023). AI and science: what 1,600 researchers think. *Nature, 621*(7980), 672–675. https://doi.org/10.1038/d41586-023-02980-0

Weber-Wulff, D., Anohina-Naumeca, A., Bjelobaba, S., Foltýnek, T., & Wawer, A. (2023). Testing of detection tools for AI-generated text. *International Journal for Educational Integrity*, *19*, Article 26. https://doi.org/10.1007/s40979-023-00146-z

Weippl, E. R., & Schrittwieser, S. (2023). Introduction to security and privacy. In H. Werthner, C. Ghezzi, J. Kramer, J. Nida-Rümelin, B. Nuseibeh, E. Prem, & A. Stanger (Eds.), *Introduction to digital humanism: A textbook* (pp. 397–414). Springer. https://doi.org/10.1007/978-3-031-45304-5_26

Werthner, H. (2024). Digital transformation, digital humanism: What needs to be done. In H. Werthner, C. Ghezzi, J. Kramer, J. Nida-Rümelin, B. Nuseibeh, E. Prem, & A. Stanger (Eds.), *Introduction to digital humanism: A textbook* (pp. 115-134). Springer. https://doi.org/10.1007/978-3-031-45304-5_18

Werthner, H., Ghezzi, C., Kramer, J., Nida-Rümelin, J., Nuseibeh, B., Prem, E., & Stanger, A. (2024). Preface. In H. Werthner, C. Ghezzi, J. Kramer, J. Nida-Rümelin, B. Nuseibeh, E. Prem, & A. Stanger (Eds.), *Introduction to digital humanism: A textbook (pp. v–x). Springer.*

Werthner, H., Prem, E., Lee, E. A., Ghezzi, C., Kramer, J., Nida-Rümelin, J., Nuseibeh, B., & Stanger, A. (2019). *The Vienna manifesto on digital humanism*. https://dighum.org/dighummanifesto/

Winter, S. (2024). The road less taken: Pathways to ethical and responsible technologies. In H. Werthner, C. Ghezzi, J. Kramer, J. Nida-Rümelin, B. Nuseibeh, E. Prem, & A. Stanger (Eds.), *Introduction to digital humanism: A textbook* (pp. 267-282). Springer.

Yeo, M. A. (2025). Academic publishing in the age of GenAI: Expanding the concept of journal impact. *RELC Journal*, *56*(2), 221-226. https://doi.org/10.1177/00336882251361159

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

Consciousness: Human, Artificial, or Both?

Viktor J. Tóth

Doctor of Philosophy (PhD), Fuller Theological Seminary, Pasadena, CA, USA

vtoth@fuller.edu

Abstract

The nature of the "intelligence" of artificial intelligence is a hotly debated notion today. Sensational claims about the successes of AI (including the declaration that synchronicity is already achieved) and warnings against its potential malice (e.g., that once it attains the level of human consciousness it will turn against humanity) are part of our everyday experience. In this article, I argue that the most extreme claims about the present state of AI are not warranted. In the first part of the article, I make the case that human intelligence/consciousness is a unique phenomenon that cannot be matched by machines. Then, in the second part, I turn to what I consider the more valid possibilities and worries about AI. I envision artificial personas (quasi-agents) working in collaboration with human agents on the betterment of the world, and counsel against the possibility of weaponizing AI. I finish the article by reflecting on some of the constructive ways for philosophical engagement in this discussion.

Original Research

Keywords:

human consciousness, artificial intelligence, quasi-agent, embodied cognition, artificial moral agents

Received: 2026/03/24; **Received in revised form:** 2026/05/16; **Accepted:** 2026/06/10; **Published online:** 2026/06/15

Tóth. V. J. . (2026). Consciousness: Human, Artificial, or Both? *Journal of Philosophical Theological Research*, (Special Issue on AI, Consciousness, and Human Agency) 28(3), 155-176. https://doi.org/10.22091/jptr.2026.17198.3657

Copyright © The Author

Introduction

"A cyborg, a robot, and a hybrid walk into a bar . . ." It sounds like the opening line of a joke, but it more or less describes the overall theme of the 2025 FX television series *Alien: Earth* (Hawley, 2025). The story purportedly takes place in 2120. As the first episode starts, a short script appears on a computer screen:

> In the future, the race for immortality will come in three guises:
> Cybernetically enhanced humans: Cyborgs
> Artificially intelligent beings: Synths
> and
> Synthetic beings downloaded with human consciousness: Hybrids
> Which technology prevails will determine what corporation rules the universe. (Hawley, 2025)

For the purpose of this article, I define these three concepts as follows: a cyborg is a human person who was born and developed naturally but receives technological enhancements that provide physical and/or mental abilities exceeding normal human capacities (Option-1[1]); a synth is a robot with a fully synthetic body and artificial intelligence (Option-2); and a hybrid is an entity exhibiting human consciousness (having been born naturally and raised in an ordinary human environment) but whose natural body dies for some reason (e.g., terminal illness), after which their consciousness is downloaded into a synthetic body (Option-3).[2] This paper aims to calrify the difference among these three options from a philosophical point of view and give an explanation of why only Option-1 is a viable prospect at this point of AI development.

Although all the above examples are taken from the world of science

1. Therefore, people who receive medical treatment as a remedy to treat disabilities (e.g., a lost limb) are not cyborgs because they do not gain superior abilities; enhancement aims to restore their natural abelites. Medical science already advanced to the level where neural interfaces can surgically placed into the cortical tissues, meaning that the person's brain is operating an artificial limb in the same way it would move organic limbs (see e.g., Ferris Jabr, "Brain Wave," *The New York Times Magazine*, "The Health Issue," May 15, 2022, 26–31, and Simon Makin, "Brain-Controlled Prosthetic Limbs Take Steps Closer to Reality," *Science News* 206, no. 9 (2024): 22–25. In popular culture both "RoboCop" and the Terminator (a.k.a. Cyberdyne Systems Model 101) are called *cyborgs* (Option-1), but in my definition only RoboCop classifies; the difference being that RoboCop was born as a human being (Alex Murphy) having human consciousness (ultimately regaining his memories), but Model 101 has artificial intelligence operating a synthetic body (even though "covered by" living tissue), thus qualifies as a *synth* (Option-2).

2. Probably the best-known example of this scenario in popular culture is Major Motoko Kusanagi from the *Ghost in the Shell* franchise. Transhumanism embraces this concept as a viable option to enhance human lifespan beyond the biological limitations of a human body.

fiction, they are generally treated as real possibilities in contemporary society. In other words, although they have not yet been realized by science, humanity feels confident that they are not only achievable but will be attained soon enough. Besides the technological complexities, this enterprise raises non-engineering-related questions arising from the fields of ethics, philosophy, theology, psychology, and economics (just to mention a few). In what follows, I will reflect on some of the philosophical aspects of the notion. I will approach the issue focusing on whether Artificial General Intelligence (hereafter, AGI)[1] is a viable option based on the capabilities current AI systems exhibit. The claim that once AGI is reached it will be inevitably followed by technological singularity makes this enterprise significant.[2]

The purpose of this article is to examine these possibilities by using the above-mentioned options as hypothetical alternatives and, in this manner, assess their achievability. First, in Sections 1 and 2, I will consider the cyborg, synth, and hybrid alternatives, using cultural paleoanthropology as a benchmark in the study of human consciousness. In Section 3, I will elaborate on the possibility of AI as an enhancement to human life. I will finish the article by pointing out some of the (in my opinion) most viable avenues for future deliberation.

The Depth and Width of Human Consciousness

In this section, I focus on Option-2 (synths) and Option-3 (hybrids). Although O-2 and O-3 are not identical options, they both operate under the assumption that "human consciousness" (I will use the abbreviation "HC" from now on) can be represented digitally as coded information flow (e.g., as an algorithm) and ultimately "run" as a program.[3] One of the auxiliary premises of this assumption is the widely accepted notion that information is something extra

1. While regular (or narrow) Artificial Intelligence (AI) is a system designed to excel in a particular task (e.g., playing chess or GO, translate between languages, or run a hospital in the most energy efficient way), AGI not only able to transfer skills form one task (or filed) to another, but also solve novel problems via leaning from the history of its interactions and thus reach the cognitive level of humans.

2. The hypothetical point in technology when a human built system transcends human comprehension.

3. I must note here that the nomenclature of the different fields of research dealing with the nature of human consciousness is perplexing; words like *consciousness*, *mind* or *soul*, *intellect*, *intelligence*, *psyche*, *brain*, or *thinking* (as a noun) are used freely and interchangeably which confuses the discussion enormously. In this work I will use the designation "human consciousness" (HC) as constantly as I can.

(or beyond) physical existence – something immaterial. This postulation then feeds into the concept "that conscious awareness can be reduced to the structural residue of information or computational process in the brain" (Frank et al., 2025, p. 15) and ultimately simulated in very sophisticated machines.

I find it somewhat startling that although the questions of what consciousness is and how consciousness works are mired in controversy (I even venture to say that the more it is studied and debated, the murkier the concept becomes), it is almost universally accepted that it is technologically reproducible. The belief that AGI is around the corner (or, according to some, already achieved) is indeed widely accepted and leads to the alarming idea that AI can somehow acquire self-consciousness and determination, and even turn against humans. This technological positivism entirely ignores the so-called hard problem of consciousness (the term was introduced by David Chalmers in 1994, but the inquiry goes back at least as far as the Presocratic philosophers in the West). The hard question concernshow organic material (i.e., animal bodies) can give rise to the phenomenon termed HC (in the idiom artificial intelligence, "artificial" basically stands for the physical part, or "hardware," and "intelligence" for HC as a hardware-based entity). Thus, in order to even entertain O-2 and O-3 as viable options, we need to take a closer look at HC.

While there is only a limited consensus about the nature of consciousness, there are some notions that most scholars agree on. One of the widely accepted opinions is that HC is the outcome of a long evolutionary process. This assumption is mainly based on the principle that if the human body is the outcome of a lengthy evolutionary process, then higher human capacities must be the outcome of the same developmental means.

For a long time, the evolutionary study of HC was instigated in a somewhat obscure way, focusing almost entirely on brain development and technological inventions, leaving other segments of human evolution neglected.[1] This – in my opinion, outdated – approach is based on a cognitivist individualistic view of human nature that endorses the notion of abstract contextualized cognition as the hallmark of rationality. As Cat Bohannon puts it so aptly: "We're terribly impressed with our brain. I'd argue we're in love with it. Which is to say, the human brain is in love with itself. If there's a single, physical trait that most scientists agree delineates humans from the other apes, it's our huge, lumpy, terribly intelligent brain" (Bohannon, 2023, p. 245). Indeed, this

1. For a thorough critique of the "cognitive approach" see Haidle, 2012. Haidle writes: "close correlation between brain size and intelligence still has to be rejected as misleading" (Haidle, 2012, p. 56). Terrence W. Deacon also criticizes "our preoccupation with brain size (Deacon, 1997, Ch. 6).

cognitivist approach implies that reason is the defining characteristic of human nature and rationality is the most important activity for human agency.[1] As an outcome of this one-sided outlook, emotions were treated as a "lower" function, a remnant of our animal ancestry. Thankfully, this attitude toward human intelligence is rapidly changing.

One of the pioneers of creating a different account of HC, one that takes emotions seriously, is Antonio Damasio. In countless publications and so many books, the Portuguese neuroscientist makes the claim that feelings and emotions are as important for the study of HC as our cognitive capacities. His groundbreaking opus, *Descartes' Error*, explains how the cognitive agenda hijacked the narrative about human consciousness from the time of Plato, and then was cemented into modern thinking via Descartes (Damasio, 1995).

The so-called "social brain hypothesis" is one of the alternative attempts to describe HC from a paleoanthropological perspective that puts emotional cognition at the center of hominin evolution. Its advocates claim that what makes humans qualitatively different from other animals is our capacity for complex and diverse social life. In short, human civilization was able to progress from mere sociality (which is found in many other species) to sociability. What distinguishes human sociality is the relative depth, the layers of affiliations and intentions, that we are able to navigate on a day-to-day basis. From this perspective, "intelligence" is not about brain size or the individual's capacity to create and use sophisticated tools, but about the ability to form multilayered, orderly societies. Humans can learn and successfully employ certain social skills that significantly enhance the sociocultural dimensions of the group. Although the "creative spark" (to utilize Agustín Fuentes's watchword in Fuentes, 2017) is indispensable, it is not the brilliant individual who maintains the possibility of human progress, but the stable, multilayered, and transgenerational social structure. In what follows, I will make some deliberations on the notions of transgenerational and multilayered societies in an attempt to relate them to AGI.

Transgenerational memory, the ability to maintain information flow between successive generations, is a must for every organism's survival. It can be described as the memory of the collective. The quality and quantity of the information stored and distributed within social systems is one of the most important marks of the intelligence of a given species. In this sense, the notion includes the ability to create decision-making processes via amplifying the residues of past experiences, and, in a similar vein, to cancel out unconstructive biases across time. However, as the above sentence indicates,

1. For a more detailed argument on this issue, see Lakoff and Johnson, 1999, pp. 3 & 408.

this "memory" is not a kind of stored knowledge in an individual's head (as passive information), but active datasets that routinely re-enter the system from a "higher" (cultural) level. During the history of human evolution, this feedback system amplified humanity's overall freedom of behavior through constant re-evaluation. In essence, this process defines the learning process of the group. This process is utterly different from the operational function of data centers, the facilities used in the maintenance and development of AGI.

Yet there is something even more special about the way living systems learn and remember. In the words of Terrence Deacon, the learning strategy of organisms is "defined by a fundamental incompleteness" (Deacon, 2006, p. 143, emphasis in original). They seem to somehow perceive what is missing or incomplete for their most desirable future (i.e., a target state). Deacon coined the term *teleodynamic* pertaining to this apparent "end-directed tendency" (Deacon, 2012, p. 271); the word comes from the Greek *telos*, which means "end" or "goal." One of the main characteristics of teleodynamic processes is the strong presence of the notion of a self that can benefit or suffer. The notion of self thus underlines the representation, functional organization, and normative characteristics of the organism. Nevertheless, this process requires several selves interacting with each other in a coordinated way. This is why the memory of an organism cannot be depicted as (passive) information, but only as a powerful medium on an always operational (group) level. This continuously active operational dynamic creates non-linear feedback loops that then amplify the initial conditions and generate progress. The dynamic, self-transcendent nature of "memory systems" is a sine qua non of HC – a component that is seemingly missing from AGI development.

The second concept I want to touch upon here in view of the evolution of HC is the notion of a layered society. This concept is closely related to the notion of individual selves (as the locus of intentionality) but extends its scope. The social brain hypothesis distinguishes six possible layers of intentionality. First-order intentionality is the awareness of my own feelings and thoughts. The second layer is my basic awareness that others around me have similar states of mind to mine. It creates the basic ability for the self to "step back" and pose the question of how and why things are as they are. On the third level, we become aware of our own states of mind, which can coexist with, but are different from, the states of mind of other individuals.[1] The fourth layer operates with the supposition that others suppose that we should coordinate

1. It is debated whether this level can exist without full-blown human language not necessarily extended beyond the level of noun-verb statements. Probably the level of Large Language Models, because it uses language without the capability the concept of what language *actually* is.

our actions in relation to an "abstract" rule (e.g., for the benefit of the group).[1] This level is achievable only via abstract representation, which presupposes language with full recursive grammar and phrase structures. A genuinely moral human community emerges only with fifth-order intentionality. On this level, one is able to handle situations in which "I believe that you suppose that I imagine that you want me to believe that" (Dunbar, 2014a, p. 58). It seems that such an abstraction requires special mental devices, such as myths and rituals.[2] Humans use allegorical and metaphorical thinking to operate on this level. What enables HC to designate "general" knowledge – knowledge that is acquired in one area to be used in other areas – is metaphorical thinking.

There is no space here to fully entertain the notion of powerful memory systems such as myths and rituals.[3] But there is no question about the significant role they played (and still play) in creating, sustaining, and changing not only cultural milieus but whole cultures. Rituals and myths are closely connected and were, without a doubt, developed in tandem. Johan Huizinga expressed it eloquently: "In myth and ritual the great instinctive forces of civilized life have their origin: law and order, commerce and profit, craft and art, poetry, wisdom, and science" (Huizinga, 1938/1950, p. 95). On the one hand, myths served (and do so even today) as collective external memory systems (Donald, 1991, p. 382), creating and maintaining the intellectual environment in which explicit consciousness develops (Peterson, 1999, p. 80). On the other hand, ritual "represents the creation of a controlled environment where the variables (i.e., accidents) of ordinary life may be displaced precisely because they are felt to be so overwhelmingly present and powerful" (Bellah, 2011, p. 135). In contemporary societies, shared social rituals and myths function as preservatives of social equilibrium amidst a seemingly corroding

1. This level is claimed to be reached by OpenAI when in July 2026 an artificial intelligent system supposedly broken free of human control. This accident is widely reported by news agencies, but I yet to find a reliable independent scholarly review of the incident. Nevertheless, note that this level is still below HC.

2. N. B.: The general population never operates on the sixth level (only people with special skills do, such as William Shakespeare. Robin Dunbar illustrates this level by explaining how Shakespeare's *Othello* works: "the audience has to *suppose* [level 5] that Iago *intends* [level 4] that Othello *believes* [level 3] that Desdemona *loves* [level 2] Cassio and that Cassio also *loves* [level 1] her (see Dunbar, "Why Only Humans Have Language," p. 440 – emphasis original). Shakespeare operates on the sixth level because he must intend that the audience believes that this is the narrative about Othello.

3. There is a supplementary notion here about the "volume" and "speed" of information processing. AI is much faster than humans when it comes to process *raw* data. However, myth in itself is a memory device that helps to process cultural information collected over countless generations. In this sense, humans are better processing this special kind of (cultural) data. My point is that to use phrases like "information" and "processing" can be misleading when it comes to understanding how *human* intelligence works.

global civilization. As I see it, AGI is inevitably a participant in this cultural milieu since it is based on the cultural setting of its programmers and trainers; however, AGI participates in such notions only as raw information. I see this as a significant difference between HC and AGI.

And the difference does not stop here. HC goes well beyond AGI by the unique way it frames memory. The human way of accumulating memory is not comprised solely of what and how we remember. It is of equal significance what we do *not* remember. Our success is equally based on our ability to ignore a large amount of information – information that we, by some means, determine to be irrelevant. We do it without running entire sets of information against another set of information. We do it so naturally, without even noticing it (for details on this issue, see Frank et al., 2025, pp. 166–172). We prioritize information based on relevance as a context-dependent framing system that "cannot be reduced to heuristic rules and symbolic representations" (Frank et al., 2025, p. 168).

So-called "second-wave" programming uses large amounts of data and attempts to analyze them by finding large-scale patterns and statistical correlations. Once technology achieves this level, AGI will (allegedly) be able to ride the wave of context dependence; however, in the process, it reproduces statistical classifications embedded in the culture of its trainers. I want to be cautious here; I recognize my limited knowledge of computer science. Nonetheless, to move forward with my argument, I propose that HC may display similar characteristics, being based on a special kind of statistical operation. Yet, to highlight what I perceive as the difference between the statistical/probability approach used by second-wave AI computation and HC, I employ here Karl Popper's notion of "propensity":

> Propensities, it is assumed, are not mere possibilities but physical realities. They are as real as forces, or fields of forces. And vice versa: forces are propensities. They are propensities for setting bodies in motion. Forces are propensities to accelerate, and fields of forces are propensities distributed over some region of space and perhaps changing continuously over the region (like distances from some given origin). Fields of forces are fields of propensities. They are real, they exist. (Popper, 1990, p. 20)

Instead of utilizing statistical (probability) calculations, HC operates within propensity fields. To rephrase, while statistical possibilities as mathematical values are successfully incorporated into AI systems, HC operates within propensities – "fields of forces" representing real causal powers. The propensity field is the environment in which the system is embedded, and it has a real impact on the outcome of the system's actions. Consequently, the

system's perspectives are based on values essential to the overall environment (as a metasystem) for its development, human culture being the most important metasystem. I venture to say that AI is twice removed from this level because the non-"value-free" dataset is provided as "valued" information through human intervention. In other words, companies that create AI (and aim to create AGI) must work around this deficiency by constantly providing the necessary "values" (via human agency) to make sure that the operation of the system does not come to a halt.[1] I will come back to these features in the next section.

All in all, I agree with Damasio that it is possible for robots (O-2) to develop consciousness, but their consciousness is probably going to be different from HC (as cited in French, 2026). However, I have no intention of ending on a dismissive note. AI is real and, just like every other powerful invention, can cause both harm and benefit in our lives. In the next section, I will consider an alternative approach to the notion.

Layers and Players: Persona as Quasi-Agent

In a recently published article, Zahra Khazaei proposes an idea that I find remarkably close to my concept of differentiating between the human self and the human persona. Khazaei proposes the concept of a "quasi-agent" (Khazaei, 2026). She defines this concept as "an entity that exhibits stable, adaptive, and often unpredictable patterns of agent-like behavior without possessing ontological subjectivity, intrinsic intentionality, or unified experiential presence" (Khazaei, 2026, p. 9). In what follows, I will link this notion to the idea of a layered society and social games.

HC is not a free-flowing information matrix separated from material existence. It is both embodied (it is revealed by and through human bodies) and embedded (it can be developed and effectively employed only within a particular culture). Most AI narratives seem to apply a mechanistic view of human existence, in which the body (or the brain) is the hardware through which consciousness (or intelligence) operates as the software. But the human body is not a machine. Properly speaking, it is not just a physical entity made up of body parts (e.g., limbs, head, torso, stomach, lungs, brain, etc.) or a composite of different small building blocks (e.g., carbon, hydrogen, calcium, etc.). It is better explained as a system of processes (e.g., circulatory system,

1. Another way to label these necessary "values" is "worldview." As long as humans are involved worldview systems are unavoidable.

digestive system, nervous system, reproductive system, etc.) manifesting as a whole within its environment (Murphy, 2018, pp. 246–1247). These subsystems are better defined by what they do than by their physical properties; they are structural constituents that might overlap with each other and might even serve different purposes over time.

I propose that human identity stems from two intertwining components based on the embedded-embodied condition: the self and the persona. On the one hand, the persona is a conglomerate of bits and pieces in our personal sociocultural toolsets, activated by the demands of playing our roles in different social settings. On the other hand, the self (or "minimal self") is the "constant background feeling of being alive … or the simple self-referential aspect of self-awareness," which "also can be imagined as a hidden, fluctuating affective relief structure; that is, the dimension of combined affective valences, emotions, dispositions, [and] habits" (Horváth, 2024, p. xvii). Thus, a self may be represented by several personas. These personas are rooted in and, ideally, managed by a unified reflective self-identity.

In the following, I use examples from the human ontogenetic process in an attempt to argue why AGI can operate as a persona (i.e., a quasi-agent) or as a placeholder for an individual entity, but not as a "full-fledged" self. Human beings do not enter life with fully developed intelligence; as newborns, we even lag behind other mammalian species. As the late German Lutheran theologian Wolfhart Pannenberg puts it, "In comparison to other higher animals human beings are born a year too soon and still in an unfinished state" (Pannenberg, 1985, p. 38).[2] On the other hand, we enter our sociocultural environment to inhabit readymade personas (i.e., roles to play in life).[3] Here, I am referring to the notion that, because most brain growth occurs after birth, the community functions as a "kind of social uterus" (Haidle & Schlaudt, 2020, p. 167).[4]

---

1. We might also describe subsystems from a biochemical point of view in which we might study the endogenous opioid system, or the neurobiological system, or the skeletomuscular system, or fascia system, as subsystems working in unison for maintaining homeostasis within the body.
2. For a wider perspective on Pannenberg's notion, see Kärkkäinen, 2015, pp. 252–55. Both Pannenberg and Kärkkäinen highlight the fact of how this specially evolved physiological trait makes "enculturation" (a psychological phenomenon) an indissoluble part of human nature.
3. For example, as newborns their *persona* allows them to cry whenever they feel like it, to expose their feces, or even to wake up at the middle of the night and require feeding when their caregiver is fast asleep. The point I am making here is that this *persona* is only a temporal entity; the time must come when it is left behind and the child must go through a metamorphosis to learn to operate different personas, ones which conforms their future circumstances.
4. Kärkkäinen uses the phrase "social womb," attributing it to Adolf Portmann, the great Swiss

There are obvious differences between "training" a human child toward intelligent behavior and "training" an AI model. For one, as the Finnish American theologian Veli-Matti Kärkkäinen points out, "Human development is facilitated rather than hindered by the deficiency of instincts, unlike other animals [and, I insert, AI]; hence, in human life a 'hiatus, a gap, between perceptions and impulses' is productive" (Kärkkäinen, 2015, p. 255) – and the opposite might be said about AI. I find only one similarity in how AI language programs and human beings are trained: both require human oversight.[1] I venture to say that this difference is even more imperative when it comes to AI development. To put it briefly, HC is an embodied and embedded process that emerged through hundreds of millions of years, and each human individual (i.e., self) is trained in their cultural environment based on their particular cultural inheritance. The training of AI, by contrast, is based on datasets provided by humans, mostly focusing on removing "noise" (partially through tokenization[2]), followed by the involvement of human "trainers" in the fine-tuning process (first for preference and then for reasoning, i.e., advanced logic). What I see here is humans (operating via human intelligence) trying to jumpstart AI by providing countless hours of labor by countless people and using a huge amount of energy (both in terms of human manpower and on the machine side).[3] And the product? Something that produces outcomes that, in some special cases, match or even exceed those of humans. Yet, it works only as an enhancement of human intelligence. Therefore, I argue that although AI can be trained by humans to occupy a quasi-agent status in our social matrix, given the current status of development, ontologically speaking, it will never become a genuine self.[4]

---

zoologist (Kärkkäinen, 2015, p. 255).

1. This assumption is based on my Google search on "how deep language programs are trained"; I used an AI generated overview, so it must be accurate!
2. This abstraction is based on the mostly outdated concepts of the structure of language by Noam Chomsky and his followers.
3. One of the worries about LLMs is that they use an enormous amount of energy, because they rely on large industrial data centers. As an example, imagine the GO competition between Lee Se-dol (South Korean former professional Go player of 9 dan) and AlphaGo (a computer program developed by Google DeepMind that mastered Go). While master Lee probably had a regular breakfast before the game (and consumes a normal number of calories per day during his lifetime, which is between 1,600–2,400 calories [= 0.00186kWh–0.002789kWh]), AlphaGo used approximately 170 kW–440 kW = 46,271,510–378,585,086 cal. during the game. It means that the magnitude of energy used by the AI is 170,000 bigger! Surly, this disparity between HC and LLMs alone must rase questions about the potential future of AI systems.
4. The *self* versus *persona* is my nomenclature for Khazaei's distinction between (human) agent and quasi-agent. My concept is explained in detail elsewhere (see Viktor J. Tóth, "The Evolution of the Human Soul: Multidisciplinary Enquiry into the Emergence of the 'Soul'" [PhD diss., Fuller Theological Seminary, 2026], "Chapters 3 and 4;" available at https://doi.org/10.13140/RG.2.2.29498.04802).

I will continue this line of reasoning in an attempt to make my argument clearer. HC, from a sociocultural perspective, is a problem-solving skill of everyday life. It is the accumulation of the kinds of social games we perform as healthy members of society. Most of the time, we do this automatically (i.e., many of our actions do not even rise to conscious awareness). More and more scholars from different fields of anthropology and brain sciences realize what some philosophers (e.g., Ludwig Wittgenstein and Maurice Merleau-Ponty, just to name the two most important) grasped almost a century ago: thinking and decision-making are not processes that take place solely in our brains. The brain, or even the whole body, is only a necessary condition for HC. In other words, HC is not a "thing" but a process – the way humans participate in a shared world.

James Wm. McClendon Jr. and James M. Smith also identify the "social matrix" as an embodied relationship (McClendon & Smith, 1975/1994, pp. 162–174). However, their explanation adds another layer to the concept. They note that a child learns language by learning language-games, in which the context and content of the words used in a given social situation are based on the situation itself (and not on dictionary forms). It is a "way of talk embodying assumptions and ways of reasoning that distinguishes the various forms" (McClendon & Smith, 1975/1994, p. 22). The conventions on which these language-games are based are not even explained to us, and they do not have to be, because "they are beneath the level of explanation, yet quietly determinative of how the language is understood, how the language-game in question is played" (McClendon & Smith, 1975/1994, p. 23). Such practices are validated by, and dependent on, the community. Thus, when Large Language Models (LLMs) are trained by humans, they are linked to this social matrix by their trainers, but without the basic concepts on which the actual language-games are formed. In other words, no matter how confidently an LLM can operate on this level, it is able to do so only because of human intervention and will probably remain in need of human supervision throughout its operation.

I think the best thing about this debate is that it helps us learn more about humanity. Marina Zilbergerts words this notion quite nicely:

> For what AI reveals with striking clarity is not the full nature of human mind but the narrowness of our definition of intelligence. Intelligence, when reduced to calculation, optimization, and symbolic manipulation, may indeed be reproducible in machines. Consciousness, however, is not. While artificial intelligence excels at replicating certain mechanical aspects of reasoning, it falters when confronted with other core capacities of the human mind, such as recognizing boundaries, responding with discernment to novel situations, or expressing (let alone experiencing) the depth and

nuance of human emotion and other qualia. (Zilbergerts, 2026, p. 135)

As a clarifying note, I point out that Zilbergerts uses *intelligence* and *consciousness* differently from my terminology. "Intelligence" for her, being reducible to data processing and optimization, is closer to my depiction of persona. In her jargon, "consciousness" is what stands for the unique way of human existence (I use both *consciousness* and *self* to highlight different aspects of this phenomenon).

In human society, being successful means being able to detect and project the right persona in each situation. Since I argued above that persona is equivalent to the quasi-agent in Khazaei's reasoning, I propose that a well-trained AI can occupy this social space. However, cultures demand that humans acquire several well-developed personas or index-identities[1], which also suggests that, within the framework adopted here, an AI persona will never be able to reach the level of HC (that is, AGI in digital networks). Being one specialized persona trained in one set of etiquette[2] as an accessory for human beings (i.e., Option-1) is plausible; however, reaching the obligatory level of changing etiquette "on the fly" based on a particular social setting (choosing from dozens of possibilities) seems to be unachievable at this point of AI development.

Natural-Born Cyborgs

The notion of enhanced humans in this article refers to people with physical, cognitive, or sensory abilities amplified beyond normal human limits using technology or genetics. The British philosopher Andy Clark made the claim decades ago that *Homo sapiens* is a "natural-born cyborg" (Clark, 2003). Clark's remark, of course, was highlighting the fact that humans evolved not just to adapt to their environment, but also to incorporate elements of their environments into their everyday lives (clothing, tools, and even animals through domestication). Not only that, but we make significant changes to our environment to make elements of it useful to us. We build and maintain well-organized societies far beyond those of every other living organism; we intuitively utilize tools and each other through seamless connections. In a way,

1. I refer to Cole's terminology here because I find it usefully for my argument (for more information on this concept see Cole, 2015).
2. I think "etiquette" here is a better choice than "protocol" since the former refers to a smaller (human) scale of operational setting (e.g., a particular social setting), while the later has an artificial (unnatural) "taste" to it.

HC can be defined by this very special way of living. HC is not a thing; it is a way of life. To use an analogy, it is not the wing that makes a bird able to fly; it is how a bird uses this special feature, adapted to the rest of its body and evolved through millions of years.

The point I make here is that our consciousness (HC) is not something that separates us from nature,[1] nor is it something that separates us from each other. Consciousness links us. Another way of expressing consciousness is living a lifestyle of constant integration. It is the integration of developmental processes within an evolutionary "framework that engages with cultural systems and processes including institutions and recognizes the fluid and entangled interfaces between individuals, groups, and community-level dynamics" (Fuentes, 2016, p. 14). We cooperate with each other and with our environment on a meta-level. In other words, "human nature is actually a meta-nature, represented by the possibility of dynamically assembling a vast gamut of extended cognitive architectures" (Wheeler & Clark, 2008, p. 3563). For us, being conscious means that we are interconnected, which means that all of our actions and capabilities, although experienced in a situated manner, are carried out in interconnection with each other. We coordinate, bond, and accomplish this so proficiently and so naturally that we do not even notice it most of the time. Following this line of thought, I see AI as an organic outcome of the evolution of HC effortlessly integrated into human consciousness.

One of the main tenets of my argument is that we must consider HC as something denoted to humankind and not as an opportunity of the individual. In other words, there is no such thing as individual consciousness; it can only be recognized and studied as a shared experience (i.e., an isolated human body would never experience what a regular human being in her social setting develops into and experiences as consciousness).[2] Thus, I see it as futile to investigate the possibility of downloading a person's consciousness into a machine. In other words, it is impossible to isolate "one" consciousness.[3] However, it is not only feasible to imagine AI becoming part of the global (or

1. The sad fact that humanity exploits nature created a false dichotomy between nature and culture. Yet, harming nature is a suicidal act, and hopefully humanity will realize this mistake soon enough.
2. As a devastative but highly informative source see "Romanian Orphanage Case" (Mackesa et al., 2020).
3. Here, I make a note that one crucial aspect of HC is not even touched upon in this article: the *un*conscious side of human intelligence. Subconscious activity was part of the investigation about human intelligence and agency from the very beginning of human thinking, but since Sigmund Freud and Carl Gustav Jung it was firmly established the into the scientific fields of psychology and psychiatry. No enquiry about HC can be adequate without weighting in the complex world of the unconscious.

local) consciousness matrix, but it is indeed happening as we speak.

Here, again, I weigh in on the notion of a layered society. Such a social construction is built on rules and regulations. Without going into details, I note here that we live our everyday lives following written and unwritten rules.[1] They are cultural artifacts that constrain our lives. Our success is very much based on how well we can adapt to them and utilize them. The current state of the AI debate is mostly about how such regulations can be implemented so that AI will continually be an asset to HC and not an adversary. I very much doubt that simply imposing laws (e.g., the "Three Laws of Robotics") will suffice here.[2] I think it is much more constructive to equip AI systems with "habitual dispositions."[3]

Mark Graves proposes that one potential way to implement the notion of habitual dispositions in AI is to train AI systems toward self-development. Such a method would assist the system in acquiring "habits" as general interpretive decision-making initiatives. Rehearsing natural encounters of the system with human culture and analyzing the possible outcomes (still with human supervision) would eventually build good habits into the system's operational repertoire and move the system toward building a stable (morally good) character. The hope is that eventually the system reaches a point where it can mediate among possible good choices in a way that is close enough to human decision-making. I propose that once this level is reached, the system achieves what Khazaei identifies as a quasi-agent, possessing all the potentials of a persona (without attaining the ontological level of self). Graves identifies "practical wisdom" as a byword for developing habits and character traits from a virtue ethics perspective for the purpose of enabling the system to consistently make morally informed, skillful decisions aligned with our ethical norms (Graves, 2026). Still, I maintain that no matter how useful the habits or practical wisdom such a quasi-agent (artificial persona) might develop, AI capabilities or apparent intelligence do not, by themselves, establish either a genuine self or genuine consciousness.

I propose that the optimal goal is not to replace human intelligence and

1. These rules are codified in civil law systems, inherited as cultural habits, encountered as institutional regulations, etc.
2. The "three Laws of Robotics" are created by the science-fiction writer Isac Asimov and introduced in his short story titled "Ruanround" (Miami, FL Analog Science Fiction and Fact, 1942). The three laws are: (1) a robot may never implement any harm on a human being; (2) a robot must always obey human orders, except if it conflict with the "First Law"; and (3) a robot must protect its own existence except if doing so would conflict with the first two laws.
3. What follows is informed by Graves, 2026, Graves, 2021, and Graves, 2025.

behavior, but to opt for kinds of intelligence that may or may not resemble human performance while being ethically responsible. In other words, the aim is not to copy the execution of an ethically satisfying action, but the outcome of it (i.e., the betterment of human existence). That is to say, AI should not copy human standards but find the best way to augment (support) human endeavors (which is not necessarily through imitation, but, at least sometimes, through the implementation of new possibilities). This objective might benefit from an approach that, instead of focusing on "rationalistic-cognitive" predispositions, treats human intelligence more as a role-playing endeavor.[1] Thus, rather than trying to create human-like agents (agents that we cannot detect as non-human), research should focus on developing quasi-agents (agents that we know are not human but are designed to help humans). Instead of trying to reproduce our intelligence as a narcissistic endeavor, why not simply focus on discovering new ways that can aid human flourishing?

I agree with Graves that achieving perfect rationality is not feasible in complicated environments where computational demands exceed any system's resources (Graves, 2025). However, it is absolutely viable to use AI as a quasi-agent to assist or substitute for humans when computational demand is high. I assess that what I reflected on above as the metaphorical and allegorical dimensions of human comprehension coincides with Graves's "motivational dimensions"; if my assessment is correct, then I also agree with him that this is the threshold that is outside the grasp of (contemporary) AI research. Nevertheless, AI can be trained, for example, to perceive human distress signals (e.g., those of a child or a dying patient in a hospital) and initiate soothing behavior. Such behavior would be indistinguishable from genuine empathy, but still qualitatively different from its human counterpart.[2]

Echoing Graves's analyses, I propose that although current self-monitoring techniques might be able to identify interpretable data within neural network representations and correlate them by identifying activation-space directions compared to moral attributes, these techniques do not yet constitute the reflexive (self-)identity required (Graves, 2021). The most significant restraint in self-monitoring AI systems is that they achieve only partial reflective self-monitoring (and never a full sense of self built upon a narrative self-perception). Graves also highlights that such constitutive identity does not emerge in isolation. The only examples we can work with are those provided by human culture and communal life, and it is very unlikely that such a

1. I do not advocate here for using "game theory" (a highly rationalistic concept) which already plays a crucial role in AI development.

2. In the same wain the fire rescue robot's detecting of higher temperature than its body can endure qualitatively defers from a human firefighter's perception of unendurable heat.

phenomenon can be reproduced in any other kind of setting.

In summary, there is currently no sufficient basis to conclude that AI is as close to gaining HC as some (mostly entrepreneurs from AI development companies) might want us to believe. Yet, there is no question about the seriousness of the enterprise when it comes to the future of human culture. AI can be a powerful instrument both as an aid to human betterment and as a destructive menace. My conclusion, therefore, is that instead of creating false anxiety about AI (or AGI) running rogue against humanity, we should attain a better understanding of how it can be used (and constrained) to aid human advancement. Here, on the one hand, we can promote "roboethics" ("robot ethics") while, on the other hand, endorsing machine ethics to develop AIs with "built-in" morality and thus create artificial moral agents (AMAs) (Boyles, 2024). AMAs are endowed with built-in, top-down ethical systems to elicit moral behavior. The aim here is not to supply AIs with preset ethical principles, but to let them develop moral behavior on their own. It is assumed – and from an evolutionary approach rightly so – that moral behavior would naturally emerge via system-environment interactions with human society (Boyles, 2024, p. 17). From an embodied and embedded perspective, I see it as a framework for a human-endowed simulation of evolutionary processes in a virtual environment.

I suppose that since AI is developed by and within human agencies, which are social environments already furnished with powerful means to deal with the aggregation of power and malicious elements, it would be "difficult for AI designers not to impose their personal convictions in addressing these issues; social choice ethics only mitigates the said impositions" (Boyles, 2024, p. 19). From this approach, it is obvious that for AMAs to function in ways that are not proper according to human ethical standards, they would have to be designed that way. This means that it is human criminal behavior that would make AI systems act against proper behavior. Thus, the "vicious cycle" of crime and punishment circles back to us. It is not an "evil AI" we need to be wary of, but "the evil inclination of the human heart" (Genesis 8:21).

Retrospect and Prospect

AI could be dangerous, like many other state-of-the-art technologies. It threatens the "average Joe" in many ways: it might eliminate jobs,[1] help terrorists create

1. See https://abcnews.com/Business/artificial-intelligence-eliminate-thousands-jobs-new-york-city/story?id=133181521. Although others who claim otherwise (see e.g., Robert

bioweapons,[1] or even drain the beauty out of mathematics.[2] In this article, I argued that one threat is false: AI, at least in its current developmental stage, is not going to gain HC. Nevertheless, AI needs to be taken seriously, and that is why I propose the following from a philosophical approach:

1. As we develop a philosophy of AI, we need to be careful to distinguish it from the philosophy of mind by taking good care of its nomenclature (I consider Khazaei's "The Challenge of Agency" a good example of this; see Khazaei, 2026). Consciousness is a notoriously elusive concept; however, in regard to AI systems, we can use the notion of "quasi-agent" and/or "artificial persona" (AP) as a background for clarifying what "human consciousness" (i.e., "full agency") means. In other words, we need to develop a new conceptual framework for this task.
2. The obvious legal responsibilities are being considered as we speak. I want to add a note of caution here, though: I would not trust big tech companies to spearhead this task by self-regulation. Yet, this is exactly what is happening, and it warrants serious concern.
3. As I argued above, AI is not a transcendent (i.e., non-material) reality. AI is hardware-dependent. And it has already reached the point where the amount of material resources (especially rare-earth elements), human labor, and energy consumption (the electricity cost of current large-scale deep-learning systems is enormous) probably does more harm to the environment (and thus to human lives) than good. In a recently published book, the authors warn: "AI is fundamentally an 'extractive industry,' one that 'depends on exploiting energy and mineral resources from the planet, cheap labor, and data at scale'" (Frank et al., 2025, p. 15).
4. Mental health issues: A recent study on human-computer interaction suggests that "users are no longer viewing AI as a mere tool but are developing meaningful emotional bonds with conversational systems" (Donard & Ribeiro, 2026). I want to add that interaction with some of these entities (i.e., LLMs) becomes part of one's everyday life history, such that any interruption in this relationship might cause distress. At this point, it seems that the information feedback loop between the system

---

Bartholomew, "AI Hysteria: Is the Sky Really Falling?", *Psychology Today* [Posted August 23, 2026], available at *https://www.psychologytoday.com/us/blog/its-catching/202608/ai-hysteria-is-the-sky-really-falling*. Bartholomew argues that it was always the case when new technologies surfaced people feared that they will take away jobs.

1. See https://screendna.org (June 2026).
2. See https://zenodo.org/records/20302944 (June 2026).

and the user has already become an emergent property (i.e., a meaningful event) in the life of the human agent. Looking back, in reminiscence of the history of other "addictive" substances (here I take the liberty of claiming that human-computer interaction has already reached the point at which the exchange itself has become a "substance" in this sense), we seem to be already too late in this matter.

5. I agree with Damasio that it is possible for robots (Option-2) to develop consciousness based on their bodies, but their consciousness is going to be different from HC (as cited in French, 2026). Therefore, I think it is very important to clarify what we mean by "artificial intelligence." Are we talking about reproducing HC, or something else entirely?

▣ Conflict of Interest

The author declares no competing interests.

References

Bellah, R. N. (2011). *Religion in human evolution: From the Paleolithic to the Axial Age*. Harvard University Press.

Bohannon, C. (2023). *Eve: How the female body drove 200 million years of human evolution*. Alfred A. Knopf.

Boyles, R. J. M. (2024). Can't bottom-up artificial moral agents make moral judgements? *Filosofija Sociologija*, 35(1), 14–22. https://doi.org/10.6001/filsoc.2024.35.1.3

Clark, A. (2003). *Natural-born cyborgs: Minds, technologies, and the future of human intelligence*. Oxford University Press.

Cole, J. (2015). Hominin language development: A new method of archaeological assessment. *Biosemiotics*, 8(1), 67–90. https://doi.org/10.1007/s12304-014-9198-8

Damasio, A. (1995). *Descartes' error: Emotion, reason and the human brain*. HarperCollins.

Deacon, T. W. (1997). *The symbolic species: The co-evolution of language and the brain*. W. W. Norton & Company.

Deacon, T. W. (2006). Emergence: The hole at the wheel's hub. In P. Clayton & P. Davies (Eds.), *The re-emergence of emergence: The emergentist hypothesis from science to religion* (pp. 111–150). Oxford University Press.

Deacon, T. W. (2012). *Incomplete nature: How mind emerged from matter*. New York: W. W. Norton & Company.

Donald, M. (1991). *Origins of the modern mind: Three stages in the evolution of culture and cognition*. Harvard University Press.

Donard, V., & Ribeiro, J. C. (2026, April). Technological mourning after AI updates: Mental health and well-being in the GPT-4o/GPT-5 transition. *Academia Mental Health & Well-Being*. https://doi.org/10.20935/MHealthWellB8256

Dunbar, R. I. M. (2014a). 'Mind the gap' or why humans aren't just apes. In R. I. M. Dunbar, C. Gamble, & J. A. J. Gowlett (Eds.), *Lucy to language: The benchmark papers* (pp. 47–60). Oxford University Press.

Dunbar, R. I. M. (2014b). Why only humans have language. In R. I. M. Dunbar, C. Gamble, & J. A. J. Gowlett (Eds.), *Lucy to language: The benchmark papers* (pp. 61–74). Oxford University Press.

Dunbar, R. I. M., Gamble, C., & Gowlett, J. A. J. (Eds.). (2014). *Lucy to language: The benchmark papers*. Oxford University Press.

Frank, A., Gleiser, M., & Thompson, E. (2025). *The blind spot: Why science cannot ignore human experience*. MIT Press.

French, K. (2026, September). Does consciousness require a body? *Nautilus*. https://nautil.us/does-consciousness-require-a-body-1284886/

Fuentes, A. (2016). The extended evolutionary synthesis, ethnography, and the human niche: Toward an integrated anthropology. *Current Anthropology*, 57(S13), S13–S26. https://doi.org/10.1086/685684

Fuentes, A. (2017). *The creative spark: How imagination made humans exceptional*. Dutton.

Graves, M. (2021). Emergent models for moral AI spirituality. *International Journal of Interactive Multimedia and Artificial Intelligence*, 7(1), 7–15. https://doi.org/10.9781/ijimai.2021.08.002

Graves, M. (2025, October 10). *Designing an AI virtuous agent* [Paper presentation]. Pacific Coast Theological Society meeting, San Francisco, CA, United States.

Graves, M. (2026). AI practical wisdom and compassion. *AI and Ethics*, 6(1), Article 39. https://doi.org/10.1007/s43681-025-00877-4

Haidle, M. N. (2012). *How to think tools? A comparison of cognitive aspects in tool behavior of animals and during human evolution* (Vol. 1, Cognitive Perspectives in Tool Behaviour). Tübingen University Press. https://tobias-lib.uni-tuebingen.de/xmlui/handle/10900/49627

Haidle, M. N., & Schlaudt, O. (2020). Where does cumulative culture begin? A plea for a sociologically informed perspective. *Biological Theory*, 15(3), 161–174. https://doi.org/10.1007/s13752-020-00351-w

Hawley, N. (Writer & Director). (2025, August 12). Neverland (Season 1, Episode 1) [TV series episode]. In N. Hawley (Executive Producer), *Alien: Earth*. FX Productions; Hulu.

Horváth, L. (2024). *The affective core self: The role of the unconscious and retroactivity in self-construction*. Springer Nature.

Huizinga, J. (1950). *Homo ludens: A study of the play elements in culture*. Beacon Press. (Original work published 1938)

Kärkkäinen, V.-M. (2015). *Creation and humanity* (Vol. 3, A Constructive Christian Theology for the Pluralistic World). William B. Eerdmans.

Khazaei, Z. (2026). The challenge of agency in the age of artificial intelligence: An ontological perspective. *Journal of Philosophical Theological Research*, 28(1), 7–14. https://doi.org/10.22091/jptr.2026.16299.3583

Lakoff, G., & Johnson, M. (1999). *Philosophy in the flesh: The embodied mind and its challenge to Western thought*. Basic Books.

Mackes, N. K., Golm, D., Sarkar, S., Kumsta, R., Rutter, M., Fairchild, G., Mehta, M. A., & Sonuga-Barke, E. J. S. (2020). Early childhood deprivation is associated with alterations in adult brain structure despite subsequent environmental enrichment. *Proceedings of the National Academy of Sciences*, 117(1), 641–649. https://doi.org/10.1073/pnas.191126411

McClendon, J. W., Jr., & Smith, J. M. (1994). *Convictions: Defusing religious relativism* (Rev. ed.). Trinity Press International. (Original work published 1975)

Murphy, N. (2018). *A philosophy of the Christian religion: For the twenty-first century*.

SPCK.

Peterson, J. B. (1999). *Maps of meaning: The architecture of belief*. Routledge.

Popper, K. R. (1990). *A world of propensities*. Thoemmes.

Wheeler, M., & Clark, A. (2008). Culture, embodiment, and genes: Unraveling the triple helix. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 363(1509), 3563–3575. https://doi.org/10.1098/rstb.2008.0135

Zilbergerts, M. (2026). Algorithms, zombies, and Maimonides' God: A Jewish theology of consciousness for the age of AI. *Tradition: A Journal of Orthodox Jewish Thought*, 58(2),122-149. https://doi.org/10.54469/R9XT5FM2

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

Narrating and Judging: Meta-Qur'anic Speech in Sūrat Ash-Shuʿarāʾ and the Limits of AI

Abdelmadjid Ben Habib

Department of Psychology, Educational Sciences & Speech-Language Pathology, Faculty of Humanities and Social Sciences, University of Abou Bekr Belkaid Tlemcen, Tlemcen, Algeria,
abdelmadjid.benhabib@univ-tlemcen.dz

Abstract

Original Research

This article examines the gap between human speech as phenomenologically intentional and language as processed by AI through statistical computation. Phenomenologically, speech originates in an intentional movement of consciousness toward meaning, interpreted within a horizon of expectation rather than simply extracted. Algorithmic processing, by contrast, proceeds through lexical frequency and co-occurrence, without an intentional subject. The article asks whether this separation is absolute, or whether a productive connection between the two is epistemologically legitimate. To sharpen the inquiry, it turns to a limit case from the Qur'an, *Sūrat Ash-Shuʿarāʾ* (Q. 26:10–68), staging intradiegetic exchanges – Moses and his Lord, Pharaoh, and the magicians – beneath a recurring extradiegetic voice, addressed to the audience rather than any character: Meta-Qur'anic Speech, comprising a narratorial register advancing the narrative and a verdictive register suspending it to render judgment. This verdictive register, in the refrain at Q. 26:67–68, renders two verdicts – condemnation for Pharaoh's people, glorification for Moses and his companions – while addressing a third, extradiegetic party: the audience. This case serves not as an exegetical end in itself, but as a philosophical instrument for testing what algorithmic processing cannot reconstruct: an act of intentional, dialogically situated judgment irreducible to lexical frequency or co-occurrence.

Keywords:

phenomenology; intentionality; algorithmic processing; dialogism; Qur'anic discourse; human agency; verdictive register; epistemology

Received: 2026/02/27; **Received in revised form:** 2026/05/03; **Accepted:** 2026/06/01; **Published online:** 2026/06/05

Ben Habib, A. (2026). Narrating and Judging: Meta-Qur'anic Speech in Sūrat Ash-Shuʿarāʾ and the Limits of AI. *Journal of Philosophical Theological Research*, (Special Issue on AI, Consciousness, and Human Agency), 28(3), 177-204. https://doi.org/10.22091/jptr.2026.17421.3685

Copyright © The Author

Introduction

What does it mean for speech to understand its meanings and to problematize its answers and echo, rather than merely process the quantity of its lexical content or recurrences? In this article, we highlight a problematic at the border of the philosophical and the theological, situated within the broader inquiry this special issue invites into questioning what is philosophically and theologically at stake in AI and human existence, in its broader sense related to identity.

The argument proceeds from an existential premise: human speech is a conscious and unconscious lived experience, structured by an intentionality at stake with each act of speech, and influenced by a historical horizon of expectation that shifts. Both the current psychosocial context and the historical context influence the construction, deconstruction, and ongoing revision of meaning. Meaning, on this view, is not simply present in an utterance awaiting extraction; it is interpreted, reconstructed, and at times acknowledged as resistant to full understanding.

Artificial intelligence approaches speech from a different epistemic register altogether: linguistic competence is achieved through the calibration of vast quantities of textual input against probabilistic and statistical models, such that output is determined by the structure of input, without problematizing the issue of why speech occurs or what its genealogy and goal are. This contrast raises three interrelated questions that structure the present inquiry.

First, is the separation between these two modes of engaging speech a genuine and irreducible one, or can a connection between them be drawn in ways that are epistemologically productive? Second, if such a connection is legitimate, under what conditions can it be established without dissolving the specificity of human linguistic experience into a numerical calculation? Third, what does this contrast reveal about the structural limits of artificial intelligence in grasping the activity of speaking as a total human subjective phenomenon, one that cannot be explained by the statistical dimension alone? To address these questions, the article turns, in its central section, to a particular revealing case: the discursive architecture of *Sūrat Ash-Shuʿarāʾ* (Q. 26:10–68), the twenty-sixth chapter of the [1]Qur'an.

This chapter stages sustained exchanges of speech between named interlocutors – a horizontal, dialogical structure in the sense elaborated by

1. All Qur'anic citations in this article follow the Saheeh International translation, accessed via Tanzil.net (https://tanzil.net). Arabic text follows the standard *rasm* ʿUthmānī as presented on the same platform.

Bakhtin – and layers over them a distinct, non-diegetic voice that narrates and arbitrates among these exchanges: adding or reporting narrative content in the first case, authorizing, correcting, or rejecting claims in the second – as when the narrated confrontation between Moses and Pharaoh's court is followed by a verdict addressed not to any character within the scene, but directly to the audience of the Qur'an: "Indeed in that is a sign, but most of them were not to be believers" (Q. 26:67).

This verdictive function is one register within a broader non-diegetic voice that recurs throughout the sura – elsewhere (e.g., Q. 26:32–33, 38, 45–46) taking a narratorial, scene-setting form rather than a judging one, a distinction developed further in the section on the discursive architecture of Sūrat Ash-Shuʿarāʾ. This distinction is analytical rather than theological, following Genette's account of narrative levels: the sura's entire reported speech is, doctrinally, divine in origin, but only certain segments break frame to address the audience directly rather than remaining within a narrated scene between characters.

This voice – here termed Meta-Qur'anic Speech – is treated in what follows not as an object of exegesis in its own right, but as a philosophical test case: an instance of language performing, in its narratorial register, an act of framing and orienting the audience's access to the events narrated, and, in its verdictive register, an act of normative, intentional arbitration between competing voices – either function irreducible to what a purely statistical treatment of the text could produce, which could only redescribe both after the fact, not perform them. The case is presented briefly and illustratively, to sharpen a philosophical argument about the limits of algorithmic language processing, not to offer an exhaustive close reading of the sura. The article proceeds as follows.

The first main section, "The Phenomenological Framework: Intentionality, Horizon, and the Hermeneutic Circle," develops the phenomenological framework – intentionality, the horizon of expectation, and hermeneutic openness – that grounds the notion of language as lived experience. The second, "Algorithmic Language Processing: What a Large Language Model Does," examines what algorithmic language processing does in practice, and why its statistical operations remain structurally distinct from understanding.

The third, "The Discursive Architecture of Sūrat Ash-Shuʿarāʾ: Problematizing Meta-Qur'anic Speech," introduces the *Sūrat Ash-Shuʿarāʾ* (Q. 26:10–68) case and problematizes the notion of Meta-Qur'anic Speech as a vertical, non-diegetic instance that both narrates and arbitrates, irreducible to horizontal dialogism or to enunciative address alone. The fourth, "Epistemological Implications: AI, Judgment, and Human Agency," draws out the epistemological implications of this case for the broader question of AI's relation to human language and identity. The Conclusion closes the article.

The Phenomenological Framework: Intentionality, Horizon, and the Hermeneutic Circle

1. Intentionality and the Eidetic Reduction

Phenomenology, since Husserl, treats consciousness as always "consciousness of something" (Husserl, 1966, p. 35): every act of awareness is directed toward an object, and this directedness – intentionality – is not an incidental feature of mental life but its defining structure. Accessing this intentional structure requires a specific methodological move. The eidetic reduction proceeds through the Epochē (Husserl, 1994; McKenna, 1997): the suspension of naive beliefs, prior knowledge, and unexamined presuppositions about the object of inquiry, in order to transform lived experience into essences – intentional structures constituted in the life of the ego, described as they give themselves to consciousness rather than as the inquirer would wish them to be. Applied to language, this means bracketing what one already assumes an utterance means, so as to describe the intentional act by which meaning is reached toward and constituted, rather than presupposing that meaning in advance. Applied to speech specifically, an utterance is not the transmission of a pre-formed content from one mind to another, but an act of consciousness reaching toward meaning. From a phenomenological perspective, "the first act of consciousness is to want to say something" (Ricœur, 2004, p. 10) – *vouloir dire* – a formulation that locates meaning not in the utterance as an inert object, but in the movement of a speaking subject toward it. To speak, on this view, is to intend: to direct oneself toward an object of meaning, to grasp something of it, and to leave something of it always still to be grasped. The listener or reader, in turn, does not extract a meaning already fully present in the words, but performs their own intentional act, reconstructing and redirecting what the speaker's utterance was reaching toward.

2. The Horizon of Expectation and the Fusion of Horizons

This intentional movement does not occur in a vacuum. Each act of speech is situated within a horizon of expectation: a historically and psychosocially shaped field within which meaning becomes interpretable at all. This horizon enters into what Gadamer calls a fusion of horizons (Gadamer, 2013). The text's own horizon of meaning is transformed in its encounter with the audience's horizon of understanding, and, reciprocally, the audience's horizon is itself reshaped by the text – an interinfluence between inherited meaning and newly constructed meaning, which reveals the collective, intersubjective dimension of the interpretive act. Meaning, on this account, is never simply present in an

utterance awaiting extraction: it is interpreted, reconstructed, deconstructed, and acknowledged as resistant to full understanding.

The horizon of expectation, and its fusion with the horizon of the interpreting subject, is thus not only what makes meaning possible; it is what makes meaning genuinely open, revisable, and contested.

3. The Hermeneutic Circle and the Erklären/Verstehen Distinction

This openness is formalized through the hermeneutic circle: interpretive knowledge of a discourse is achieved through a continuous movement between the general structure of the discourse and its particular constitutive utterances, and back again. Ricœur and Gadamer insist that this circularity is the very shape that understanding takes. Central to this tradition is the distinction, inherited from Dilthey (1947) and developed by Ricœur (1977), between *Erklären* (explanation) and *Verstehen* (interpretation and understanding). Explanation proceeds by tracking the linguistic traces of a discourse – its phonological, syntactic, lexical-semantic, and pragmatic-intentional features – to reconstruct its partial, local meanings. Interpretation and understanding, by contrast, are performed by the reader, who intentionally binds together the meanings that explanation has surfaced, aiming at a global synthesis. Crucially, Ricœur does not treat these as opposed operations: the global construction of meaning is built from the partial constructions that explanation supplies, and the partial constructions of explanation are themselves oriented, from the start, by the anticipation of a global meaning. Explanation and understanding dialectically influence each other; they are not sequential stages one can perform without the other.

4. Performativity: Speech as Act

This dynamic is not only semantic; it also has a performative dimension. Discourse does not only say something – it does something to its audience. It promises, commands, questions, exhorts, consoles, and addresses; the audience, in turn, responds through its own speech acts – prayer, contestation, assertion, interpretation – extending and transforming the field of meaning in a continuing exchange. This is precisely the interaction between saying and doing that speech act theory (Austin, 1962; Searle, 1985; Searle & Vanderveken, 1985) is designed to capture: an utterance is analyzed for what it does – the illocutionary act it performs, the illocutionary point that act is defined by, and the perlocutionary effect it elicits in its audience.

5. Axiological and Practical Intentionality

Husserlian intentionality is not limited to theoretical aim. In the Fifth Logical Investigation, Husserl distinguishes objectifying acts – representation, judgment, which aim at an object as simply given – from non-objectifying acts (*Gemütsakte*), always founded upon the former: "Each intentional experience is either an objectifying act or has its basis in such an act" (Husserl, 2001, p. 167). Among these non-objectifying acts, two families can be distinguished: acts of valuation (*Wertungsakte*) – joy, sorrow – which require a full affirmation of their object, and conative acts – the wish – which are satisfied by a mere neutral modification. This double distinction establishes, from within the phenomenological apparatus itself, an axiological/evaluative dimension and a practical/conative dimension of intentionality, which the later subsection "The Addressee's Axiological and Practical Capacity" will mobilize to characterize what the reception of a verdictive act requires of its addressee.

6. Dialogism as a Structural Correlate of Intentionality

This phenomenological and performative account of meaning has a linguistic correlate in Bakhtin's notion of dialogism (Bakhtin, 1981, 1984b, 1986): no utterance is self-contained. Every utterance stands in relation to utterances that precede it, anticipates the utterances that will respond to it, and is oriented, implicitly or explicitly, toward another voice, whether that voice is present, absent, historical, or imagined. Dialogism is not incidental to intentionality; it is its social and discursive expression. The movement of consciousness toward meaning, the horizon within which that meaning becomes interpretable, and the performative force of an utterance are never the operations of an isolated subject, but are always already oriented toward, and shaped by other speaking subjects and their utterances. In brief, dialogism embodies the echo of meaning.

7. Implications for What Follows

Together, these elements – intentionality accessed through the eidetic reduction, the horizon of expectation and its fusion with the interpreting subject's own horizon, the hermeneutic circle and the *Erklären*/*Verstehen* dialectic, the performative force of speech as act, and dialogism as the social form intentionality takes in discourse – constitute the phenomenological dimension against which the next section examines what algorithmic language processing does, and does not, do. The question that follows is not whether a statistical system can track linguistic traces of some kind – clearly, it can – but whether

tracking traces of this kind can coincide with the intentional, horizon-bound, performative, and dialogically oriented activity of wanting to say something.

Algorithmic Language Processing: What a Large Language Model Does

1. From Intentional Meaning to Statistical Structure

The phenomenological account developed in the preceding section locates meaning in an intentional act – a consciousness reaching toward an object, bracketed and described through the Epochē. Meaning is also taken up in context, situated within a horizon that fuses with the horizon of another; here appears the encounter between the text or discourse and its audience. Against this account, a large language model (LLM) operates on an entirely different epistemic register. An LLM is a program trained by accumulating vast quantities of textual data and calibrating a probabilistic model over that data, such that its output is determined by the statistical structure of its input: patterns of lexical frequency, occurrence, and co-occurrence across enormous corpora (Harris, 1954; Lenci & Sahlgren, 2023). What such a system discovers is what manifests itself in a text at the level of distributional regularity – which words tend to appear near which others, how often, and in what configurations – not what a speaking subject intended by producing that text.

This is not a claim that statistical regularity is unrelated to meaning. Distributional patterns are not arbitrary; they are, in an important sense, the sedimented trace of countless intentional acts of speech performed by human speakers over time – a trace an LLM accesses only in an algorithmic register. In other words, a trace of an intentional act is not itself an intentional act. An LLM's output is the product of a mechanical process of pattern-completion over these traces, not the product of a consciousness bracketing its assumptions, reaching toward an object, and constituting meaning through that reaching.

2. No Epochē, No Bracketing

The eidetic reduction described in "Intentionality and the Eidetic Reduction" requires a subject capable of suspending its own naive assumptions in order to describe an object as it gives itself to consciousness. An LLM has no assumptions to suspend, in the phenomenological sense, because it has no consciousness directed toward an object in the first place: what is sometimes described, loosely, as a model's "bias" is not a naive presupposition available for bracketing, but a

statistical artifact arising from its training distribution – a structural property of the data, not a phenomenological standpoint that a subject occupies and can step back from by applying the Epochē. Where the phenomenologist's Epochē is an act performed by a subject upon its own presuppositions, an LLM's output is shaped by its accumulated training data, with no analogous act of stepping back available, since only consciousness can take this kind of distance from itself.

3. No Horizon, No Fusion

The subsection "The Horizon of Expectation and the Fusion of Horizons" described meaning as situated within a horizon of expectation that fuses, dialogically, with the horizon of an interpreting subject. An LLM does not occupy a horizon in this sense. Its "context window" is sometimes described in language that borrows from hermeneutics (the model "attends to" prior context), but this is a technical metaphor, not a phenomenological one: attention, in the technical sense used in these architectures, is a weighted statistical operation over token representations (Vaswani et al., 2017), each token representation itself being a sedimented function of that token's occurrence and co-occurrence across the training corpus – not an intentional orientation of a subject toward an object within a historically shaped field of expectation. There is no fusion, because there is no second horizon on the model's side to fuse with the text's; there is only the text's statistical structure being processed against weights derived from other texts.

4. Explanation Without Understanding

The *Erklären*/*Verstehen* distinction offers perhaps the most precise way to locate what an LLM does and does not do. What such a system performs is, in a strict sense, a radicalized and automated form of *Erklären*: the tracking of linguistic traces at a scale and speed no human explanatory procedure could match – which is why LLMs can be genuinely useful tools for surfacing patterns in a corpus. But *Verstehen*, on Ricœur's account, is not simply more explanation performed faster and at greater scale – it is a distinct operation, performed by an interpreting subject who binds partial meanings into a global qualitative synthesis through an intentional act oriented by anticipation of that very synthesis. An LLM does not anticipate a global meaning it is reaching toward – a meaning that, in the phenomenological sense, is subjectively given and non-quantifiable. This is compatible with, though distinct from, the point made from within computational linguistics itself: that a system trained only on linguistic form has no principled path to meaning, since the mapping from form to

communicative intent is underdetermined by text alone (Bender & Koller, 2020).

An LLM computes the statistically most probable continuation of a sequence, given patterns observed in prior data. The dialectical bond Ricœur insists upon – explanation feeding understanding, understanding orienting explanation – has, on the model's side, only one term: there is accumulation and calibration, but no reciprocal movement toward a global, intended sense that is reflected upon, because there is no speaking subject for whom that sense is being constituted.

5. Simulated Performativity

The subsection "Performativity: Speech as Act" described speech as performative: an utterance does not only say something, it does something. It promises, commands, questions, and elicits a responsive act from its audience. An LLM can produce output with the surface grammatical form of a promise, a command, or a question. What it cannot do is perform the illocutionary act in Searle's sense, because illocutionary force depends on a speaker's intention and on uptake by another subject within a shared framework of convention – conditions an LLM's output can imitate at the level of form without satisfying them at the level of the act itself. The output can be redescribed, after the fact, as resembling a promise or a verdict; it cannot, on this account, constitute one, for want of being intentionally preconceived by a speaking subject, in the particularity of that subject's intention.

6. Simulated Dialogism

The subsection "Dialogism as a Structural Correlate of Intentionality" described dialogism as intentionality's social expression. An LLM's output can be made to exhibit surface features of dialogic orientation – referring back to prior turns, adjusting register in apparent response to a prompt – because such patterns are statistically present in the human-generated text the model was trained on. This is, in essence, Firth's distributional dictum operating at the level of dialogic form: "you shall know a word by the company it keeps!" (Firth, 1957, p. 11) – except here, the "word" is an entire conversational register, and the "company" is the accumulated statistical pattern of countless human dialogues in the training corpus.

What is absent is the underlying intentional orientation of one consciousness toward another that, on the phenomenological account, is what dialogism actually consists in. The system redescribes the surface shape of

dialogism; it does not enact the intentional relation dialogism names – a relation in which, as Bakhtin insists, even a single voice already echoes multiple other voices, unpredictably, from within.

7. Convergence and Divergence with Recent Philosophical Literature

Three recent studies examine, from a related angle, the question of LLM intentionality and linguistic agency. Gubelmann (2024) draws on the Kantian opposition between the autonomous organism and the heteronomous mechanism to argue that speaking – describing, ordering, lying – presupposes an intention arising from within a self-governing agent; an LLM, a heteronomous mechanism, merely executes a purpose specified from outside by its designer, so that it processes language without ever speaking in the full sense. Attah (2025), for his part, contests that the absence of Gricean communicative intention – the strong requirement that a speaker intend the hearer to recognize that very intention – suffices to disqualify LLMs: once this requirement is relaxed, nothing prevents, on his account, an LLM from producing illocutionary effects. Attah nonetheless explicitly leaves open the question of perlocutionary effects – the real, embodied transformation of an audience's conduct or understanding – which the LLM's disembodied condition cannot produce. Skiba (2026), finally, generalizes the no-assertion argument into a no-speech-act argument altogether: it is not that LLMs lack the norms governing such an act (Grice's Cooperative Principle would in principle apply to any speaker) – it is that they lack the capacity to be genuinely subject to them, for want of being able to care about a sanction, what he terms attention-deduction.

The present article extends this last line of argument, while shifting its ground: rather than examining the LLM's capacity to produce an ordinary speech act, it examines its capacity to occupy either end of a specific verdictive act – the position from which it is produced, and the position to which it is addressed (see the subsections "The Addressee's Axiological and Practical Capacity" and "Why Redescription Is Not Ruling"). It is precisely on this register that the limit noted by Attah becomes decisive: the verdictive of Q. 26:67–68 does not merely solicit illocutionary recognition of its content, but a real, perlocutionary transformation of the audience it addresses – exactly the register in which the LLM's disembodied condition is structurally lacking.

8. What Remains to Be Tested

The claims of this section are, so far, general and structural. To test them concretely, the article turns, in the next section, to a specific textual case in which an act of

intentional arbitration between competing voices – not merely the tracking or redescription of lexical patterns – is performed within the text itself: the discursive architecture of the Qur'anic chapter *Sūrat Ash-Shuʿarāʾ* (Q. 26:10–68).

The Discursive Architecture of Sūrat Ash-Shuʿarāʾ: Problematizing Meta-Qur'anic Speech

1. Corpus and Method

The passage under examination (Q. 26:10–68) stages a dense sequence of speech events, alternating between exchanges internal to the narrated action and a recurring voice addressed to the audience of the Qur'an. Table 1 presents the full sequence as a working inventory; the discussion that follows treats only a small number of these instances closely, in keeping with the illustrative, non-exhaustive scope announced in the introduction.

Table 1. Speech Events in *Q. 26:10–68*

Verses	Interlocutors	Register
10 (frame)	God's speech to the Qur'anic audience	narratorial
10–17 (quoted)	Moses and his Lord	horizontal (asymmetric)
18–24	Moses and Pharaoh	horizontal
25	Pharaoh and his court	horizontal
26	Moses, Pharaoh, and his court	horizontal
27	Pharaoh and his court	horizontal
28	Moses, Pharaoh, and his court	horizontal
29–31	Pharaoh and Moses	horizontal
32–33	God's speech to the Qur'anic audience	narratorial
34–37	Pharaoh and the eminent of his court	horizontal
38	God's speech to the Qur'anic audience	narratorial

Verses	Interlocutors	Register
41–42	The magicians and Pharaoh	horizontal
43–44	Moses and the magicians	horizontal
45–46	God's speech to the Qur'anic audience	narratorial
47–51	The magicians and Pharaoh	horizontal
52 (frame)	God's speech to the Qur'anic audience	narratorial
52 (quoted)	God and Moses	horizontal
53	God's speech to the Qur'anic audience	narratorial
54–56	Pharaoh and his people	horizontal
57–60	God's speech to the Qur'anic audience	narratorial
61 (frame)	God's speech to the Qur'anic audience	narratorial
61–62 (quoted)	Moses's companions and Moses	horizontal
63 (frame)	God's speech to the Qur'anic audience	narratorial
63 (quoted)	God and Moses	horizontal
63(cont.)–66	God's speech to the Qur'anic audience	narratorial
67–68	God's speech to the Qur'anic audience	verdictive

2. Meta-Qur'anic Speech: Level, Function, and Act

Before turning to the horizontal exchanges and the recurring non-diegetic voice in detail, it is necessary to specify, in three analytical steps, what kind of phenomenon Meta-Qur'anic Speech names: its narrative level, its narrative function, and the type of speech act it performs. First, at the level of narrative structure. Genette's account of narrative levels distinguishes the extradiegetic, comprising "the narrative of events as taken on by the narrator's discourse," from the intradiegetic, comprising "the narrative of events as taken on by a character's discourse" (Genette, 1988, p. 62). This criterion is discursive rather than biographical: what determines a given utterance's level is not the identity,

authority, or power of the speaker, but which discursive role that utterance performs at the moment it occurs. This distinction resolves a difficulty the sura's own structure poses directly. God speaks in this passage in two capacities that must be kept analytically separate.

At Q. 26:10–17, God addresses Moses within a bounded narrated scene – issuing a directive, receiving Moses's reply – and here God's speech is performed as a character's discourse: however commanding its content, it remains intradiegetic, on a level with the sura's other horizontal exchanges (Moses and Pharaoh, etc.). Elsewhere in the same sura (Q. 26:32–33, 38, 45–46, 67–68), a formally distinct voice addresses not any character but the audience of the Qur'an directly, situated outside any narrated scene: here the same underlying source of speech is performing the narrator's discourse, and is accordingly extradiegetic. It is only this second capacity – narrator's discourse, addressed to the audience rather than to a character – that the present article designates Meta-Qur'anic Speech. Two terminological equivalences follow and are used consistently throughout the article. First, horizontal dialogism, in the Bakhtinian sense employed as reciprocal exchange between named characters, is, in Genette's terms, always intradiegetic: every horizontal exchange examined in this article is character's discourse, situated within a bounded narrated scene. Second, a narratorial frame – a brief clause introducing a horizontal exchange within a single verse – is not a third category, but an instance of the same narratorial register of Meta-Qur'anic Speech examined at Q. 26:32–33, 38, and 45–46 in "The Narratorial Register: Expositive Acts": narrator's discourse performing the narrating function, here compressed into a scene-setting clause immediately preceding a quoted intradiegetic exchange.

Several verses combine both discursive roles within a single verse: a narratorial frame reporting that an event occurred, immediately followed by a quoted utterance belonging to a horizontal exchange. At Q. 26:10, "And [mention] when your Lord called Moses" is narrator's discourse, extradiegetic; the quoted command that follows, "[Saying], 'Go to the wrongdoing people'," is character's discourse, intradiegetic, belonging to the horizontal exchange developed through Q. 26:17. The same structure recurs at Q. 26:61: "And when the two companies saw one another" is the extradiegetic narratorial frame, while "The companions of Moses said, 'Indeed, we are to be overtaken!'" is the intradiegetic quoted exchange continuing into Q. 26:62.

In both cases, Genette's discursive-role criterion allows the two levels to be separated cleanly within a single verse. Second, at the level of narrative function. Not every extradiegetic utterance in the sura performs the same work. Narratology distinguishes a narrator's function of narration proper –

recounting events – from a narrator's function of commentary, which Rimmon-Kenan characterizes as operating "either on the story or on the narration" (Rimmon-Kenan, 1983, p. 98), encompassing interpretation, judgement, and generalization directed at the narratee. Meta-Qur'anic Speech, as it recurs across Q. 26:10–68, performs both functions at different points. The instances at Q. 26:10 (frame), 32–33, 38, 45–46, and 61 (frame) exemplify the narrating function: each reports that an event took place without evaluating it. The instance at Q. 26:67–68 exemplifies the commentary function: rather than advancing the narrated action, it steps back to interpret and pass judgment on what has already been recounted. Third, at the level of speech act. It is specifically in its commentary function that Meta-Qur'anic Speech performs what Austin's taxonomy of illocutionary acts identifies as a verdictive: the act of delivering a finding, ruling, or reckoning on a matter of fact or value, characteristic of "a jury, arbitrator, or umpire" (Austin, 1962, p. 150).

This verdictive classification excludes three competing candidates that likewise carry evaluative content without being verdictives in Austin's technical sense. A bare assertion with evaluative content ("Pharaoh is a tyrant") would describe a state of affairs tinged with moral judgment, without the institutional gesture of rendering a verdict on a case already heard. A behabitive – expressing indignation, blaming – would be an affective reaction to another's conduct, not a finding rendered from a position of arbitration external to the dispute. An expositive – situating an utterance within the unfolding discourse – would structure the narrative without rendering any judgment on its substance. The refrain at Q. 26:67–68 escapes all three: it intervenes after a complete sequence of narrated events, as a finding rendered on a closed case; it renders two opposed verdicts simultaneously – condemnation and glorification — the binary structure characteristic of a judgment rather than an isolated remark; and it is addressed to a third, extradiegetic party, the audience, rather than to the parties themselves, much as a court's verdict, while concerning the parties to a case, is delivered before a public that stands outside the dispute itself yet is directly present to hear and heed it. The refrain at Q. 26:67–68 – "Indeed in that is a sign, but most of them were not to be believers. And indeed, your Lord – He is the Exalted in Might, the Merciful" – does not narrate a further event; it renders a finding on the response just narrated and affirms a claim about the nature of God, precisely the kind of act Austin's category names. The narratorial instances, by contrast, are closer to what Austin terms expositive: situating an event within the ongoing discourse, without rendering judgment upon it. On this basis, Meta-Qur'anic Speech names, throughout this article, an extradiegetic voice that recurs in Q. 26:10–68 in two distinct registers: a narratorial register, performing the narrating function through expositive acts, and a verdictive register, performing the commentary function through the Austinian verdictive

act. Both registers are extradiegetic; the horizontal exchanges among Moses, Pharaoh, the magicians, and Moses's companions remain, throughout, intradiegetic – character's discourse, whether reciprocal, as in the Moses/Pharaoh confrontation examined in "Horizontal Dialogism: Moses and Pharaoh," or internally asymmetric, as in the Moses/his Lord exchange (Q. 26:10–17) developed further in "Internally Asymmetric Horizontal Dialogism: Moses and His Lord."

3. The Narratorial Register: Expositive Acts

The narratorial instances of Meta-Qur'anic Speech – Q. 26:32–33, 38, 45–46 – share a common structure: each reports an event in the unfolding confrontation between Moses and Pharaoh's court without commenting on its significance. "So he threw his staff, and suddenly it was a serpent, manifest. And he drew out his hand; thereupon it was white for the observers" (Q. 26:32–33). This verse reports the two signs Moses performs before Pharaoh's court – the staff transformed, the hand made luminous – in the flattest possible narrative register. There is no evaluative vocabulary here, no address to the audience beyond the bare act of narration, and no ruling on how the event ought to be received. In Austin's terms, this is closer to an expositive act: it situates an event within the unfolding discourse, making "plain how our utterances fit into the course of an argument or conversation" (Austin, 1962, p. 151). The verse's function is to advance the narrated action – to tell the audience what happened – not to tell the audience what to think or what to say. The same structure recurs at Q. 26:38 – "So the magicians were assembled for the appointment of a well-known day" – a pure scene-transition, again without evaluative content; and at Q. 26:45–46 – "So Moses threw his staff, and at once it devoured what they falsified. So the magicians fell down in prostration" – which reports the outcome of the contest but does not yet pass judgment on it; the judgment on belief and unbelief is reserved for the sura's later verdictive instance. This deferral is itself significant: it shows that the narratorial register is not simply commentary that happens not to evaluate, but a distinct discursive operation, confined to advancing and situating the narrated sequence, leaving evaluation to a formally separate moment that prepares the narratee to receive, in due course, the judgment whose sense emerges precisely from the narration just assimilated.

4. The Verdictive Register: The Refrain as Verdictive Act

The verdictive register is instantiated by the refrain at Q. 26:67–68: "Indeed in that is a sign, but most of them were not to be believers. And indeed, your Lord

– He is the Exalted in Might, the Merciful." Unlike the narratorial instances, this utterance does not advance the narrated sequence but instead turns back upon the sequence just narrated to render a finding on it. Two distinct findings are delivered in immediate succession, both addressed directly to the audience rather than to any character: first, a ruling on the epistemic and moral standing of "most of them" – Pharaoh's people – with respect to the signs just recounted; second, a ruling on the nature of God, affirmed in the same verdictive breath. This is precisely the structure Austin's category of the verdictive was designed to capture: "Verdictives consist in the delivering of a finding, official or unofficial, upon evidence or reasons as to value or fact" (Austin, 1962, p. 152).

The Meta-Qur'anic voice here occupies exactly this structural position – not a participant in the narrated dispute, but an instance standing outside the dispute, ruling on it after the fact, for an audience that was never a party to the original exchange at all. Two features of this verdictive act are worth isolating. First, the verdict is rendered retrospectively: it presupposes the entire preceding sequence of intradiegetic exchanges as the evidentiary basis on which the finding is delivered – "indeed in that," referring back to the sign just narrated. Second, the verdict is doubly addressed: it speaks about Pharaoh's people in the third person while speaking to the Qur'anic audience in the implied second person, positioning that audience as the party for whom the finding is actually intended. Pharaoh's people are judged; the Qur'anic audience is the one instructed by the judgment. This double structure raises the question of whether such an act could, even in principle, be reconstructed by a purely statistical treatment of language. This question is taken up directly in the section "Epistemological Implications: AI, Judgment, and Human Agency."

5. The Addressee's Axiological and Practical Capacity

The verdictive register at vv. 67–68 does not merely close the narrative file that precedes it: it interpellates its addressee. The judgment rendered bears on an *āyah*, a "sign," whose mode of reception the Qur'an itself designates elsewhere: "[…] that they might reflect (*yaddabbarū*) upon its verses (*āyātihi*)" (Q. 38:29). The term *tadabbur* – whose Form V verbal pattern carries a nuance of progressivity, and whose root (د - ب - ر, "to track") designates the movement by which one pursues, beyond the surface appearances of words, *ẓawāhir al-alfāẓ*, their concealed meanings, *al-maʿānī al-maknūna* (Ibn ʿĀshūr, 1984, vol. 23, p. 252) – does not designate passive registration, but an active interpretive uptake.

This requirement exceeds the mere "uptake" that Austin's account requires of any ordinary illocutionary act: the verdictive solicits, from its addressee, an

active axiological and practical capacity, deployed, as the subsection "Axiological and Practical Intentionality" showed, on an evaluative plane (weighing the merit of the judgment rendered) and a practical plane (orienting oneself, in consequence, toward a course of action or a belief of one's own). It is precisely this double capacity that the LLM lacks – not through technical insufficiency, but because its reception of utterances remains, by construction, a statistical treatment of the *ẓawāhir al-alfāẓ*, without meditation oriented toward a judgment of value of its own. The LLM can thus occupy neither the position from which the verdictive is produced (see "Meta-Qur'anic Speech: Level, Function, and Act"), nor the position to which it is addressed.

6. Summary of the Distinction

Table 2 summarizes the contrast developed across the subsections "The Narratorial Register: Expositive Acts" and "The Verdictive Register: The Refrain as Verdictive Act."

Table 2. Summary of the Narratorial/Verdictive Distinction

(Q. 26:10–68, Meta-Qur'anic Speech)

	Narratorial register (e.g., 32–33, 38, 45–46)	**Verdictive register** (67–68)
Narrative function	Narration (advancing the sequence)	Commentary (evaluating the sequence)
Austinian act	Expositive	Verdictive
Addressee	Qur'anic audience (implicit)	Qur'anic audience (explicit, doubly structured)
Relation to prior narrated events	Continues them	Renders judgment retrospectively upon them

7. Horizontal Dialogism: Moses and Pharaoh

If Q. 26:10–17 exhibits horizontal dialogism marked by internal asymmetry, the confrontation between Moses and Pharaoh at Q. 26:18–31 exhibits horizontal dialogism in its more fully reciprocal form: two named interlocutors,

neither subordinate to the other in address, contesting a claim through successive, mutually responsive speech acts – precisely the structure Bakhtin's dialogism anticipates, in which "every utterance… is oriented toward a response" (Bakhtin, 1978, p. 103, my translation; see "Dialogism as a Structural Correlate of Intentionality"). The exchange opens with Pharaoh recalling Moses's upbringing at court and his past offense: "Did we not raise you among us as a child, and you remained among us for years of your life? And [then] you did your deed which you did, and you were of the ungrateful" (Q. 26:18–19). This is an assertive, in Searle's sense, but its illocutionary force is that of an accusation, establishing Pharaoh's authority over the terms of the exchange before any theological content is raised. Moses's reply neither rejects the factual claim nor submits to its accusatory force: "I did it, then, while I was of those astray. So I fled from you when I feared you. Then my Lord granted me wisdom and prophethood and appointed me [as one] of the messengers" (Q. 26:20–21). Moses concedes the fact, reframes its significance, and pivots the exchange onto new ground – his own status as messenger. From a Bakhtinian perspective, this reframing enacts what Bakhtin calls the microdialogue, in which "all words in it are double-voiced, and in each of them a conflict of voices takes place" (Bakhtin, 1984a, p. 74). Moses's utterance takes up the very grammatical form of Pharaoh's accusation and redirects it: Pharaoh – "…and you were of the ungrateful" (Q. 26:19); Moses – "…while I was of those astray" (Q. 26:20).

The parallel construction – "you were of [X]" answered by "I was of [Y]" – retains Pharaoh's syntactic frame while substituting error for the moral failing Pharaoh's accusation implied, before Moses turns the accusation back onto Pharaoh: "And is this a favor of which you remind me – that you have enslaved the Children of Israel?" (Q. 26:22). This is dialogism in its adversarial register: each utterance answers the one before it by reframing the terms on which the exchange continues. The confrontation's theological center follows when Pharaoh asks, "And what is the Lord of the worlds?" (Q. 26:23) – a request for definition that Moses answers, across three successive exchanges, with an expanding series of assertives: "The Lord of the heavens and earth and that between them" (Q. 26:24); "Your Lord and the Lord of your first forefathers" (Q. 26:26); "Lord of the east and the west and that between them, if you were to reason" (Q. 26:28). Pharaoh does not engage these assertives on their own terms. Instead, at two points, he turns away from Moses to address a third party within the narrated scene: "Do you not hear?" (Q. 26:25), addressed to "those around him," and "Indeed, your 'messenger' who has been sent to you is mad" (Q. 26:27), again addressed to his court rather than to Moses.

This resembles the structure of Meta-Qur'anic Speech examined above – an utterance stepping outside the immediate exchange to address a wider audience –

while remaining, in fact, entirely intradiegetic: Pharaoh's audience is his own court, a party internal to the narrated scene, not the audience of the Qur'an. The contrast sharpens, rather than blurs, the criterion established in "Meta-Qur'anic Speech: Level, Function, and Act": what distinguishes Meta-Qur'anic Speech is not the rhetorical gesture of turning to address a third party, but the specific identity of that addressee – the Qur'anic audience, standing entirely outside the diegesis, as opposed to a character positioned within it. The exchange concludes with a shift in illocutionary type. Having failed to answer Moses's assertives on their own terms, Pharaoh issues a commissive – a threat committing himself to future action: "If you take a god other than me, I will surely place you among those imprisoned" (Q. 26:29). Moses's reply is a challenge: "Even if I brought you proof manifest?" (Q. 26:30), to which Pharaoh responds, "Then bring it, if you should be of the truthful" (Q. 26:31) – setting up the sign-performance narrated in the sura's subsequent verses. The confrontation thus moves, across its span, through assertive, accusatory, definitional, commissive, and challenge-based speech acts, each responsive to the one preceding it – a dense concentration of reciprocal illocutionary activity that stands in contrast to the single directive-and-request structure of Q. 26:10–17 examined in the next subsection. This structural density is not incidental: both confrontations stage a contest between divergent frameworks of authority, Moses's monotheistic claim set against the rival framework shared by Pharaoh and the magicians, and it is precisely this divergence that generates challenge as their organizing illocutionary stake. The exchange at Q. 26:10–17, examined in the next subsection, occurs within a single shared framework – Moses never contests the legitimacy or identity of his interlocutor – and is, correspondingly, devoid of challenge.

8. Internally Asymmetric Horizontal Dialogism: Moses and His Lord

Verses Q. 26:10–17 present a horizontal exchange, intradiegetic in Genette's sense (Genette, 1988): God addresses Moses, and Moses replies, within a single bounded narrated scene, distinct from the recurring extradiegetic voice examined above under Meta-Qur'anic Speech. Yet this exchange is not symmetrical in the way the sura's other horizontal exchanges are. Its speech-act structure marks a hierarchy internal to the dialogue itself, prior to and independent of any subsequent Meta-Qur'anic arbitration. The exchange opens with a directive: God commands Moses, "Go to the wrongdoing people" (Q. 26:10), specifying the mission in the verses that follow – "the people of Pharaoh. Will they not fear Allah?" (Q. 26:11). Moses's reply does not simply accept the command; it responds with an objection grounded in fear of the consequences: "My Lord, indeed I fear that they will deny me, and that my breast will tighten

and my tongue will not be fluent, so send for Aaron. And they have upon me a [claim due to] sin, so I fear that they will kill me" (Q. 26:12–14).

God's opening utterance is a directive in Austin's and Searle's shared sense; Moses's reply is a request, seeking accommodation – the sending of Aaron – rather than a counter-command. This asymmetry is marked lexically as well as illocutionarily. Moses addresses God as *rabbī* – "my Lord" – a possessive form of address that positions the speaker as subordinate to the one addressed. Where God's utterance moves downward, as an order directed at Moses, Moses's utterance moves upward, as a request directed at his Lord. The exchange is horizontal in Genette's narratological sense – both parties are characters within a single narrated scene – but it is not symmetrical in Bakhtin's dialogic sense: the two utterances stand as a directive and the request it occasions, marked by an asymmetry of address that recurs nowhere else in this form among the sura's other horizontal exchanges.

This observation carries a methodological consequence. It would be possible to mistake this asymmetry – God's authority, Moses's subordination, the directive-then-request structure – for an instance of Meta-Qur'anic Speech, on the grounds that God's utterance here orchestrates the mission that structures the rest of the narrated action. The subsection "Meta-Qur'anic Speech: Level, Function, and Act" addressed this possibility directly: by Genette's discursive-role criterion, this exchange remains intradiegetic, since God's utterance here is performed as a character's discourse, addressed to another character within a bounded scene, and not as the narrator's discourse addressed to the Qur'anic audience. What Q. 26:10–17 demonstrates instead is that horizontal dialogism, in this sura, is not uniformly symmetrical: internal hierarchy and directive/request asymmetry can structure an intradiegetic exchange without that exchange thereby becoming extradiegetic. This refinement – internally asymmetric horizontal dialogism – stands as a finding in its own right, distinct from, though compatible with, the narratorial/verdictive distinction developed for Meta-Qur'anic Speech above.

9. Synthesis

The analysis developed in this section establishes a three-level distinction. At the level of narrative structure, Genette's discursive-role criterion separates the intradiegetic exchanges among the sura's characters from the extradiegetic voice addressed to the Qur'anic audience (see "Meta-Qur'anic Speech: Level, Function, and Act"). At the level of narrative function, this voice divides into a narratorial register, advancing the sequence without evaluating it, and a verdictive register, suspending it to render judgment (see "The Narratorial

Register" and "The Verdictive Register" above). At the level of speech act, this functional distinction maps onto Austin's own categories: the narratorial register performs expositive acts; the verdictive register performs the act his taxonomy names, exactly, verdictive (see the same two subsections).

The verdictive act at Q. 26:67–68 presupposes the entire preceding intradiegetic sequence as its evidentiary basis and renders its finding retrospectively: two intradiegetic verdicts – condemnation for Pharaoh's people, glorification for Moses and his companions – while instructing a third, extradiegetic party, the Qur'anic audience, invited to identify with both outcomes. This double structure is legible even in the verse's own vocabulary, where the two divine attributes conjoined – *al-ʿazīz* ("the Almighty," the attribute of rigor answering the condemnation of Pharaoh's people) and *al-raḥīm* ("the Merciful," the attribute of grace answering the glorification of Moses and his companions) – correspond respectively to the fate of the two parties. This invited identification is itself doubly named – once by Ricœur's theoretical category of narrative identity, the constitution of a subject's self-understanding through exemplary narratives (Ricœur, 1990), and once by the Qur'an's own internal term, *ʿibra*, the "lesson" a narrative offers "for whoever fears [Allah]" (Q. 79:26), applied by the Qur'an itself to the Moses/Pharaoh narrative. This question is taken up directly in the section "Epistemological Implications: AI, Judgment, and Human Agency."

Epistemological Implications: AI, Judgment, and Human Agency

1. Returning to the Question

The preceding section closed by asking whether the verdictive act performed by Meta-Qur'anic Speech at Q. 26:67–68 – a retrospective, doubly addressed ruling on a preceding sequence of intradiegetic exchange – could, even in principle, be reconstructed by a purely statistical treatment of language. The apparatus developed in the sections on the phenomenological framework and on algorithmic language processing supplies a direct answer, though one that requires distinguishing between what algorithmic language processing can redescribe and what it can perform. An LLM could, without difficulty, produce a string of text resembling Q. 26:67–68: given sufficient training data containing Qur'anic exegesis and commentary on the sura's refrain structure, a model could generate an utterance with the same propositional content, following the same recurring pattern identified across the sura. This is not in dispute: surface reproduction of the verdictive utterance's form is well within the reach of a system operating on lexical frequency, occurrence, and co-

occurrence, particularly given how often this refrain itself recurs across the Qur'anic corpus on which such a model would likely be trained. What such a system could not do is perform the act Q. 26:67–68 performs, for exactly the reasons that account specifies at each of its five points.

2. Why Redescription Is Not Ruling

The transition from the intentional apparatus built in the phenomenological and algorithmic sections to the conclusion regarding the LLM's verdictive incapacity holds in a single chain: producing a verdictive act requires occupying an arbiter's position, intentionally oriented toward an audience whose horizon of understanding one seeks to transform (see "The Phenomenological Framework: Intentionality, Horizon, and the Hermeneutic Circle") – and it is precisely this intentional structure, not mere conversational passivity, that the LLM constitutively lacks (see "Algorithmic Language Processing: What a Large Language Model Does"). Symmetrically, on the addressee's side (see "The Addressee's Axiological and Practical Capacity"), the LLM cannot position itself axiologically and practically before the judgment rendered either – a lack that echoes, by extension rather than direct citation, the structure of Skiba's (2026) argument on the speaker: where Skiba shows that the LLM cannot care about its normative standing as a potentially sanctioned speaker, the foregoing analysis shows, symmetrically, that it cannot care about its normative standing as a potentially addressed recipient of a verdict.

First, an Austinian verdictive presupposes a speaker positioned to render judgment – occupying the adjudicating role of a figure such as an umpire or magistrate (Austin, 1962) – a standpoint from which the preceding sequence of events can be evaluated as evidence. As the subsection "Synthesis" showed of Q. 26:67–68's own double verdict, this standpoint is not merely adjudicative, but sovereign: capable of condemning one party and affirming or glorifying another from a position of authority external to the dispute itself. The subsection "No Epochē, No Bracketing" established that an LLM has no standpoint of this kind to occupy: it cannot bracket presuppositions through an Epochē, because it has no naive presuppositions belonging to a consciousness in the first place, only statistical regularities belonging to a training distribution. A verdictive act requires someone who judges; an LLM's output, however verdictive in surface form, has no such someone behind it. A further consequence follows from Rimmon-Kenan's account of narrator reliability. The very possibility of judging a narrator's evaluations as reliable or unreliable – "Contrasts and incongruities in the narrator's language alert us to a possible unreliability in the narrator's evaluations, though not necessarily in his reporting of facts" (Rimmon-Kenan, 1983,

p. 102) – presupposes a speaking subject whose evaluations could diverge from the facts being reported: a gap between what is said and what is meant that only a genuine evaluative standpoint makes possible. An LLM's output admits of no such gap: its statistically generated continuations are neither reliable nor unreliable evaluations, because they are not evaluations at all.

Second, the verdict at Q. 26:67–68 renders two intradiegetic verdicts – condemnation for Pharaoh's people, glorification for Moses and his companions – while instructing a third, extradiegetic party, the Qur'anic audience: a structure that presupposes exactly the fusion of horizons described in "The Horizon of Expectation and the Fusion of Horizons" and shown, in "No Horizon, No Fusion," to have no correlate on the model's side. An LLM's "attention" to a preceding context is a weighted statistical operation over token representations; it is not an orientation toward an audience whose own horizon of understanding the utterance seeks to reshape.

Third, the retrospective character of the verdict – its dependence on the entire preceding sequence of intradiegetic exchange as evidentiary basis – is an act of *Verstehen* in Ricœur's sense: a global synthesis binding together partial meanings into a single judgment. The subsection "Explanation Without Understanding" established that an LLM performs, at most, a radicalized *Erklären* – tracking traces at scale – without the reciprocal movement toward a global, intended sense that *Verstehen* requires, because there is no speaking subject for whom that sense is being constituted.

Fourth, the verdictive act is, definitionally, a performative act in Austin's original sense – delivering a verdict does not describe a state of affairs so much as bring a ruling into force. The subsection "Simulated Performativity" established that an LLM's output can imitate the surface grammatical form of such acts without satisfying the felicity conditions on which their illocutionary force depends. An output resembling a verdict is not thereby a verdict.

Fifth, and finally, the verdictive act operates within precisely the dialogic structure examined in "Simulated Dialogism": it presupposes, and responds to, the entire preceding sequence of horizontal voices as a real dialogic history. An LLM's apparent responsiveness to context reproduces the form dialogic orientation takes in its training data, without the underlying intentional orientation of one consciousness toward another that dialogism actually consists in.

3. A Qualified Answer to the Article's Opening Questions

The analysis conducted from the phenomenological framework through "Why Redescription Is Not Ruling" supports a qualified, not an absolute, answer to the

question of the separation between intentional and algorithmic language. This separation is not absolute at the level of redescription: an LLM can track, reproduce, and recombine the lexical and distributional traces that a verdictive act like Q. 26:67–68 leaves in a corpus – precisely what makes such systems useful for surfacing patterns across a body of text. But it is absolute at the level of the act itself: no accumulation of traces, however extensive, constitutes the performing of a verdictive – an act that requires a subject positioned to judge, oriented toward an audience whose horizon it seeks to reshape, and situated within a real dialogic history. An LLM can process the residue of judgment; it cannot judge.

4. Implications for Human Agency and AI

This asymmetry bears on the broader question this issue poses concerning AI, consciousness, and human agency. If the capacity to render a verdictive act – occupying a standpoint from which prior events become evidence for a judgment addressed to another's horizon of understanding – is constitutively unavailable to a system operating by accumulation and calibration alone, then this capacity marks a precise, textually argued boundary of human linguistic agency that algorithmic language processing does not cross. The case examined is a limit case, but the structural point it isolates – the difference between tracing a judgment's residue and performing the judgment itself – is not confined to Qur'anic discourse: it applies wherever human speech renders a verdict, though rarely as legibly. The Meta-Qur'anic Speech case demonstrates this point with particular clarity: rather than merely happening to instantiate the distinction cleanly, it stands as an exemplary case in the eidetic sense set out in "Intentionality and the Eidetic Reduction" – one in which the formal boundary between narrating and judging is made visible as such, where in ordinary speech the same boundary is instantiated only diffusely.

Conclusion

This article set out from a phenomenological premise: that human speech is structured by an intentionality irreducible to the statistical treatment of linguistic traces. *Sūrat Ash-Shuʿarāʾ* (Q. 26:10–68) served as a test case: beneath a dense sequence of intradiegetic exchanges, a recurring extradiegetic voice – Meta-Qur'anic Speech – divides into a narratorial register that advances the narrated sequence and a verdictive register that suspends it to render judgment. This last register, instantiated in the refrain at Q. 26:67–68, was shown to perform what Austin's taxonomy names a verdictive act: a finding rendered retrospectively,

on the evidentiary basis of the entire preceding exchange, doubly addressed – two intradiegetic verdicts, condemnation for Pharaoh's people, glorification for Moses and his companions – while instructing a third, extradiegetic party, the Qur'anic audience.

The article's central claim is that such an act cannot, even in principle, be reconstructed by a system operating through the accumulation and calibration of lexical frequency, occurrence, and co-occurrence – not because such a system is technically limited in ways future engineering might overcome, but because the act requires precisely what the preceding sections specified an intentional, judging, dialogically situated subject to possess, and a statistical model, by its own constitutive design, to lack: a standpoint from which prior events become evidence, an orientation toward an audience whose horizon of understanding the utterance seeks to reshape, and a real dialogic history within which the judgment intervenes. An algorithmic system can trace the sedimented residue these acts leave in a corpus; it cannot perform the acts themselves.

The case examined here is deliberately narrow – a single, illustrative passage, without claim to exhaustive coverage of Meta-Qur'anic Speech across the wider Qur'anic corpus – and its extension beyond Qur'anic discourse, to courts, criticism, or everyday moral appraisal, remains a plausible hypothesis rather than a completed demonstration. What the analysis establishes is a qualified answer to the question with which the article began: the gap between lived, intentional human speech and its algorithmic redescription is not uniform across every register of speech. It is at its narrowest where language merely reports, and at its widest, and most consequential, precisely where language judges.

▣ Conflict of Interest

The author declares no competing interests.

References

Attah, N. O. (2025). Do language models lack communicative intentions? *Synthese*, *205*, 187. https://doi.org/10.1007/s11229-025-05022-6

Austin, J. L. (1962). *How to do things with words*. Oxford University Press.

Bakhtin, M. (1978). *Esthétique et théorie du roman* (D. Olivier, Trans.). Gallimard.

Bakhtin, M. (1981). *The dialogic imagination: Four essays* (M. Holquist, Ed.; C. Emerson & M. Holquist, Trans.). University of Texas Press.

Bakhtin, M. (1984a). *Problems of Dostoevsky's poetics* (C. Emerson, Ed. & Trans.). University of Minnesota Press.

Bakhtin, M. (1986). *Speech genres and other late essays* (C. Emerson & M. Holquist, Eds.; V. W. McGee, Trans.). University of Texas Press.

Bakhtin, M. (1984b). Les genres du discours. In *Esthétique de la création verbale* (A. Aucouturier, Trans.; pp. 263–308). Gallimard.

Bender, E. M., & Koller, A. (2020). Climbing towards NLU: On meaning, form, and understanding in the age of data. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5185–5198. https://doi.org/10.18653/v1/2020.acl-main.463

Dilthey, W. (1947). *Le monde de l'esprit* (M. Rémy, Trans.). Aubier Montaigne.

Firth, J. R. (1957). A synopsis of linguistic theory, 1930–1955. In *Studies in linguistic analysis (Special volume of the Philological Society)* (pp. 1–32). Blackwell.

Gadamer, H.-G. (2013). *Truth and method* (J. Weinsheimer & D. G. Marshall, Trans.). Bloomsbury Academic.

Genette, G. (1988). *Narrative discourse revisited* (J. E. Lewin, Trans.). Cornell University Press.

Gubelmann, R. (2024). Large language models, agency, and why speech acts are beyond them (for now) – A Kantian-cum-pragmatist case. *Philosophy & Technology*, *37*, 32, 1–24. https://doi.org/10.1007/s13347-024-00696-1

Harris, Z. (1954). Distributional structure. *Word*, *10*(2–3), 146–162. https://doi.org/10.1080/00437956.1954.11659520

Husserl, E. (1966). *Méditations cartésiennes: Introduction à la phénoménologie* (G. Peiffer & E. Levinas, Trans.). Vrin.

Husserl, E. (1994). *L'idée de la phénoménologie: Cinq leçons* (A. Lowit, Trans.). Presses Universitaires de France.

Husserl, E. (2001). *Logical investigations* (Vol. II; J. N. Findlay, Trans.). Routledge. (Original work published 1900–1901; English translation first published 1970.)

Ibn ʿĀshūr, M. (1984). *Tafsīr al-Taḥrīr wa-al-Tanwīr* (Vol. 23). al-Dār al-Tūnisiyya li-l-Nashr.

Lenci, A., & Sahlgren, M. (2023). *Distributional semantics*. Cambridge University Press.

McKenna, W. R. (1997). Epochē and reduction. In L. Embree, E. A. Behnke, D. Carr, J. C. Evans, J. Huertas-Jourda, J. J. Kockelmans, W. R. McKenna, A. Mickunas, J. N. Mohanty, T. M. Seebohm, & R. M. Zaner (Eds.), *Encyclopedia of phenomenology* (pp. 177–180). Kluwer Academic Publishers.

Ricœur, P. (1977). Expliquer et comprendre : Sur quelques connexions remarquables entre la théorie du texte, la théorie de l'action et la théorie de l'histoire. *Revue Philosophique de Louvain*, *75*(25), 126–147. https://doi.org/10.3406/phlou.1977.5924

Ricœur, P. (1990). *Soi-même comme un autre*. Seuil.

Ricœur, P. (2004). *À l'école de la phénoménologie*. Librairie Philosophique J. Vrin.

Rimmon-Kenan, S. (1983). *Narrative fiction: Contemporary poetics*. Routledge.

Saheeh International. (1997). *The Qur'an: English meanings*. Al-Muntada Al-Islami Trust.

Searle, J. R. (1985). *Expression and meaning: Studies in the theory of speech acts*. Cambridge University Press.

Searle, J. R., & Vanderveken, D. (1985). *Foundations of illocutionary logic*. Cambridge University Press.

Skiba, L. (2026). Artificial speech acts. *Analysis*, *86*(2), anag022. https://doi.org/10.1093/analys/anag022

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*, 5998–6008. https://papers.nips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

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

A Critical Evaluation of the Fitting-Recipient Argument for the AI Ensoulment Hypothesis

Reza Dargahifar

Ph.D. in Philosophy of Mind, Assistant Professor, Faculty of Strategic Sciences and Technologies, University of Qom, Qom, Iran dargahifarreza@gmail.com

Abstract

Will advanced artificial intelligence systems one day become ensouled? An affirmative answer to this question, which is an important issue in the philosophy of artificial intelligence, has various metaphysical, ethical, legal, social, and political implications. One of Brian Cutter's two arguments in his article "The AI Ensoulment Hypothesis" in defense of the possibility that the AI ensoulment hypothesis might be realized in artificial general intelligence is the Fitting-Recipient Argument. In this article, after presenting some preliminary remarks on AGI, Cutter's argument is explained together with its underlying assumptions and key concepts, and his reasons for the truth of its two premises are formulated and analyzed. It is then shown that the first premise of the argument, which Cutter treats as expressing a sufficient condition, is in fact defensible only as a necessary condition. Accordingly, the argument commits the fallacy of affirming the consequent, that is, treating a necessary condition as a sufficient one, and therefore fails to establish its conclusion. Consequently, ensoulment cannot be inferred from fittingness for ensoulment, and thus the Fitting-Recipient Argument cannot provide a basis for Cutter's intended conclusion that we should take seriously the possibility of artificial general intelligence becoming ensouled.

Original Research

Keywords:

artificial general intelligence (AGI); AI ensoulment; fitting-recipient argument; substance dualism; philosophy of artificial intelligence; philosophy of mind; affirming the consequent

Received: 2026/02/29; **Received in revised form:** 2026/05/18; **Accepted:** 2026/06/12; **Published online:** 2026/06/18

Dargahifar, R. (2026). A Critical Evaluation of the Fitting-Recipient Argument for AI Ensoulment Hypothesis. *Journal of Philosophical Theological Research*, (Special Issue on AI, Consciousness, and Human Agency), 28(3), 205-232. https://doi.org/10.22091/jptr.2026.17187.3655

Copyright © The Author

Introduction

The first episode of the educational English-language television series *How Do You Do* (a Canadian production from 1988) begins with Frankie Stone, the creator of a humanoid robot, tightening the final screw and then pressing the power switch. The robot opens its eyes and, after seeing Frankie and hearing his greeting, begins to respond with some difficulty and hesitation. When Frankie asks how it is feeling, it says, "I am …" and, without finishing the sentence, asks, "Who am I?" Frankie replies that its name is CHIPS (an acronym for Cybernetic Heuristic Integrated Perception System).

With the emergence of successful AI technologies and the projection of various scenarios concerning their future development, one of the questions that has been and continues to be raised in the philosophy of AI is whether an artificially intelligent system might one day genuinely possess free will, rather than merely simulate voluntary behavior. Could it actually understand and know, and be aware of itself and its environment, rather than merely behave in ways that would justify others in attributing consciousness, understanding, and cognition to it? More fundamentally, could it one day possess a soul or mind? If so, could it be regarded as a living being and as a new species in the natural world? For many people, particularly some AI engineers, such questions are meaningless or pointless, and it seems highly unlikely, if not impossible, that an intelligent system could genuinely possess a soul or mind. Yet some views and theories concerning what it is for something to be alive or to possess a soul entail, at least, the possibility that life, soul, and mind could be realized in certain advanced intelligent systems.

We human beings ordinarily take it for granted that we possess a soul, or at least that we are conscious, and we can also accept that animals – and sometimes even plants – possess a soul or are conscious. By contrast, the idea that a machine could possess a soul or be conscious strikes us as highly implausible; at least for the time being, it is difficult for us to accept, or even regard as possible, such a prospect. In *The Mysterious Flame*, Colin McGinn recounts a dialogue from a novel by Terry Bisson between an alien explorer and his commander, who have come to observe the Earth. The two are astonished that terrestrial beings – human beings – are made of meat and yet are sentient and conscious, and that they have succeeded in constructing machines capable of transmitting signals. For them, it was difficult to accept the possibility that a creature made of meat could be conscious, dream, and love, or that a brain made of meat could think (where "meat" is used as a generic term for the material constituting the human body) (McGinn 1999, 7–8).

Oberg (2023) offers a significant reconceptualization of the "AI ensoulment"

hypothesis by systematically decoupling the theological concept of the soul from the psychological concept of the self. He argues that the soul, as presented in Abrahamic scriptural traditions, is necessarily tied to transcendence and postmortem continuity, whereas the self is fully determinative in purely earthly, material terms and is therefore potentially applicable to artificial systems. Drawing on and revising Kristjánsson's (2010) self-theory, Oberg defines the core Self as the sum of self-defining traits, self-directing and self-evaluating traits, consciousness, and bodily presence (Self = Set 1 + Sets 2 & 3 + C + Bp). He further grounds this model in a two-tiered, materialist account of cognition in which pre-aware, intuitive, and emotional processes dominate, and consciousness is explained as complex neural representation rather than as evidence of anything immaterial. On this basis, he argues that an artificial Self – though almost certainly different from a human Self – is not metaphysically impossible, provided that consciousness can be instantiated non-organically through representational and associative systems analogous to neural "maps" and "tags." Oberg is careful to distinguish this from full personhood, noting that an AI Self might lack the embeddedness and social identity characteristic of human persons, which would reduce but not eliminate the ethical stakes. He concludes by calling for an open and ongoing public dialogue about the moral and legal status of potentially self-aware artificial systems, while explicitly declining to answer whether such systems could meaningfully claim a "soul" for themselves. Oberg does not argue that AI can or should be "ensouled" in any religious or metaphysical sense. Rather, he systematically separates the concept of "soul" from the concept of "self" and then asks whether an artificial self (not soul) could be constructed – a condition he calls being "en-solved" (a deliberate neologism, though the pun on "ensouled" is clearly intended). His conclusion is cautiously affirmative: an artificial Self is conceivable if consciousness can be understood materialistically, but everything hinges on how we define consciousness and whether it can arise non-organically.

Tiwari and Tomar (2026) advance a categorical rejection of the AI ensoulment hypothesis, arguing that artificial intelligence can never possess a soul or consciousness – not merely as a technological limitation but as a philosophical and metaphysical impossibility. Drawing on both Western and Indian philosophical traditions, they contend that AI remains confined to language and computation, whereas human consciousness is grounded in lived, embodied, experiential subjectivity that cannot be simulated or engineered. From the Western canon, they marshal Descartes' dualism, Husserl's intentionality, Nagel's subjective character of experience, Chalmers' hard problem, and Searle's Chinese Room to demonstrate that consciousness entails features – self-reflection, intentionality, qualia, semantic understanding – that resist reduction to computation. Indian traditions, they argue, are even more radical in treating consciousness as ontologically prior to matter: Advaita

Vedanta identifies ātman with brahman; Sāṃkhya distinguishes puruṣa from prakṛti; Jainism separates jīva from ajīva; Kashmir Śaivism emphasizes vimarśa (reflexive self-awareness); and Bhartṛhari grounds meaning in śabda-brahman and sphoṭa. Pairing Wittgenstein's language games with Bhartṛhari's holistic semantics, the authors conclude that AI processes linguistic fragments without inhabiting the forms of life or experiential depth that give language meaning. They critically rebut functionalism, transhumanism, emergentism (including Integrated Information Theory), and ethical arguments for AI personhood, maintaining that simulation is never equivalent to subjectivity. Their conclusion is unequivocal: AI's boundaries are linguistic and computational, while human consciousness transcends language through lived experience, making the notion of a soulful AI both metaphysically impossible and philosophically incoherent. This position stands in direct contrast to Oberg's (2023) materialist account, which, while also rejecting the theological soul, argues that an artificial Self is conceptually achievable if consciousness is understood as materially instantiable representational processing. The tension between these two positions – Tiwari et al.'s ontological impossibility thesis and Oberg's cautious materialist optimism – frames the central debate in the literature on AI consciousness and soul.

Swann (2021) approaches the question of AI ensoulment from a Christian exegetical and theological perspective, arguing that AI cannot possess a soul in the biblical sense. Focusing on the Hebrew concept of *nephesh* (soul) as developed in Genesis, Swann contends that the human soul is not a detachable, immaterial substance but the conjunction of a physical body and the God-given breath of life, resulting in a unified, integrated being who is a soul rather than merely having one. The human *nephesh* is further distinguished by its divine mandate to serve as the Image of God and curate creation – a functional identity that AI, as a human tool, lacks entirely. Swann argues that a digital consciousness is inherently non-corporeal and dichotomous (program plus housing), and therefore cannot constitute the integrated body–breath unity that defines a biblical soul. He further notes that sentience does not entail possession of a soul, since spiritual beings possess consciousness but are not considered souls. While Swann concludes that AI cannot be ensouled, he reframes the ethical question: the real theological concern is not whether AI has a soul but whether human souls are prepared to take responsibility for their autonomous creations. AI, unlike a nuclear weapon, has determinative power, and the burden of moral accountability falls entirely on its human creators. This position stands as a distinct theological-exegetical contribution to the debate, contrasting with Oberg's (2023) materialist account of the Self and Tiwari et al.'s (2025) philosophical impossibility thesis, and together these three perspectives illustrate the range of contemporary positions on AI consciousness and soul.

In an article entitled "The AI Ensoulment Hypothesis," Brian Cutter argues

that some future AI systems will acquire an immaterial soul (Cutter 2025, p. 1). He proposes the Artificial General Intelligence Ensoulment Hypothesis and offers two arguments in support of it.

In what follows, after making some preliminary remarks about artificial general intelligence, I will first explain the AI Ensoulment Hypothesis as formulated by Cutter and then reconstruct the Fitting-Recipient Argument by distinguishing its premises and explaining and defending them. Finally, I will criticize the argument by rejecting its first premise. It is worth noting that Bálint Békefi (2026) has also criticized this argument, but he raises an objection to its second premise. Thus, my strategy for criticizing the argument differs from Békefi's.

It is also worth noting that Tijanovic's "The Virtual Ensoulment Hypothesis: Can Virtual Individuals Be Ensouled?" (forthcoming) extends Brian Cutter's (2025) AI ensoulment hypothesis from artificial intelligence to individuals inhabiting virtual worlds. Working explicitly within a substance-dualist framework, Tijanovic argues by analogy that if human beings can be ensouled, virtual individuals may likewise be ensouled insofar as they are functionally equivalent to real individuals in three respects emphasized in Cutter's framework: access to environmental inputs, a sufficiently complex behavioral repertoire, and inferential capacities partly grounded in memory. Tijanovic then introduces a novel objection, the "argument from value," according to which virtual worlds may lack values necessary for ensoulment – causal stability, truthfulness, and structural stability. He argues that the first two do not ultimately preclude virtual ensoulment, whereas the structural instability of virtual worlds constitutes a more serious challenge, since their dependence on computers makes their continued existence vulnerable to external interruption or termination. Thus, while Cutter primarily argues for the possibility of ensoulment in AI on the basis of functional capacities, Tijanovic extends this functionalist strategy to virtual individuals and tests its limits by introducing environmental and axiological conditions that may bear on fitness for ensoulment. Whereas Cutter's central concern is whether an artificial physical system can be sufficiently fit to receive an immaterial soul, Tijanovic asks whether functionally comparable individuals in computer-generated worlds can satisfy the broader conditions required for ensoulment. He ultimately remains relatively agnostic about virtual ensoulment, while maintaining that the possibility should remain open.

Artificial General Intelligence (AGI)

Artificial General Intelligence (AGI), which some refer to as strong AI or use

the term to designate something similar, is a term for AI systems that perform a variety of cognitive activities, particularly thinking, at a human level or at a level beyond and more advanced than that of human beings. Such systems perform a wide range of tasks, much as humans do. They are said to actually think, rather than merely behave as though they think. Some claim that, if such systems are constructed, they will possess sensation and consciousness, much like human beings (Cycleback 2019, 11; Leon with ChatGPT 2025, 242–243).

The AGI project aims at intelligence as a whole – an intelligence with multiple dimensions that can be deployed across a wide variety of situations. Once fully implemented, an AGI system will very likely resemble the human brain/mind in a number of respects (Goertzel & Wang 2007, 1–2). The project and the pursuit of AGI have been subject to various criticisms, including the claims that its realization is impossible; that there is no such thing as general intelligence; that it is still too early to work on it; that AGI is nothing more than hype; that research on AGI is not worthwhile because it is too difficult to be successfully realized; and, indeed, that AGI is dangerous (Goertzel & Wang 2007, 4–8).

Approaches to Creating Artificial General Intelligence

One approach that may appear to provide a reliable route to achieving AGI is Emulating the Human Brain, that is, creating an exact replica of the human brain in a digital simulation, from the highest level of organization down to the atomic level. In this sense, creating AGI is merely an engineering problem. Given that general intelligence, in the sense in which it is exhibited by human beings – who are particular configurations of atoms – is possible, all that is required is to analyze these configurations and atomic arrangements in complete detail and then reconstruct them in a computer. AGI would then emerge and come into existence through the application of a particular form of nanotechnology and within *in-silico* physics and computational space. According to this approach, the human brain serves as the guide and driving force for the production of AGI (Goertzel & Pennachin 2007, 17).

On this approach, if the appropriate physical organization of matter (at the atomic level) is reproduced, mind and intelligence will emerge as a consequence. It is important that the precise material composition of the system does not play an exclusive or determinative role; rather, what matters is its arrangement, configuration, and functional organization. According to this view, mindedness is not dependent on a biological substrate, and the brain is merely one possible substrate; silicon can fulfill the same role. If matter can be organized precisely at the atomic scale, intelligence will emerge. On this

view, to engineer the mind, one must engineer matter. The mental world is an extension of the material world, provided that matter possesses the requisite organization. This kind of view is based on the belief in the multiple realizability of the mind and intelligence and, likewise, on the principle of substrate independence, or substrate invariance, according to which mind and intelligence can be realized independently of their material substrate.

Another approach has been termed Emulating the Human Mind, that is, constructing virtual neurons, virtual synapses, and virtual activations of these components. According to this approach, the overall structure of the human brain is emulated and replicated, but not the details in which that structure is implemented and realized. The aim here is to emulate the human mind as it is studied by cognitive psychologists, while the manner in which the human mind is realized in the human brain is considered entirely irrelevant. The traditional logic-based approach to artificial intelligence belongs to this camp (Goertzel & Pennachin 2007, 19–20). As will be discussed and clarified below, Cutter appears to have advanced his account of AGI ensoulment on the basis of an approach more closely resembling Emulating the Human Mind.

Another approach is to achieve AGI by creating artificial life (see Goertzel & Pennachin 2007, 22–25), while yet another is the Integrative Approach (see Goertzel & Pennachin 2007, 26, 75–76).

The AI Ensoulment Hypothesis

As noted above, Brian Cutter argues for the claim that some future AI systems will acquire immaterial souls. Accordingly, his claim is not that: (a) current intelligent systems possess souls; (b) all future AI systems will become ensouled; or (c) the soul is made of matter or is a product of computational complexity. Rather, his claims are that: (α) a material system can become ensouled; that is, ensoulment is embodied, and an immaterial soul is in some way connected with a material, physical body; (β) the soul is immaterial, that is, it is not made of physical matter; (γ) an artificial system can be ensouled, and ensoulment is not exclusive to naturally occurring biological systems; rather, more generally, any material system possessing certain properties and capacities – whether natural or artificial, terrestrial or extraterrestrial – can be ensouled; and (δ) ensoulment is related to intelligence (whether intelligence is a prerequisite for ensoulment or intelligence is a manifestation or consequence of ensoulment).

Cutter appears to mean that a material system is capable of receiving a soul when it is capable of being intelligent; that is, intelligence is a prerequisite for ensoulment. If, however, intelligence itself is a consequence of ensoulment,

then we must consider what the necessary and sufficient conditions for the ensoulment of a material system are; in other words, under what conditions does a material system acquire an immaterial soul?

It is already clear from this that Cutter's argument rests on two premises, as he himself explicitly states: (1) that substance dualism is correct with regard to the relation between soul and body, since, according to this theory, an ensouled being possesses both a material body (i.e., a material system suitably constituted and appropriately disposed to receive a soul) and an immaterial soul; and (2) that a material system must possess some particular property or aptitude for ensoulment, since it is evident that not every material system is ensouled. Hence, possession of a soul depends on the fulfillment of certain conditions in the material entity. Cutter assumes that an Artificial General Intelligence (AGI) system – which has not yet been achieved but may be developed in the future – will possess the requisite suitability and aptitude for ensoulment (Cutter 2025, 1).

Basic Assumptions

Cutter begins his discussion with the assumption that substance dualism is correct in the case of human beings. It seems clear, however, that his assumption should not be that this theory is correct only of human beings. Rather, his claim and hypothesis require the theory to be correct of any material ensouled being. Nevertheless, the starting point of his discussion is the truth of substance dualism in the case of human beings. In other words, he begins with human ensoulment and human beings' status as substances composed of two distinct kinds of substance, and seeks to extend ensoulment to certain other material systems. On this basis, every living human body is associated with an immaterial entity that, in some way, constitutes the basis or fundamental bearer of distinctively human mental capacities, such as consciousness, thought, and rational agency. The fundamental question, then, is: given this assumption, should we expect future AI systems to possess souls as human beings do? Could highly advanced future artificial intelligences also come to possess souls?

He calls this hypothesis the AI Ensoulment Hypothesis and, by means of two arguments offered in its support, seeks to establish that this hypothesis should at least be taken seriously. Given that we human beings have souls, he argues, we should not be certain that computers will not be ensouled at all. He maintains that the hypothesis should not simply be dismissed as a fantasy or a remote possibility. In his view, if we accept the existence of souls in the case of human beings, we cannot, without sufficient argument, be confident that future artificial beings will never possess souls (Cutter 2025, 1). Cutter's

argument is a conditional: if substance dualism is true, can we then say with confidence that AI will never possess a soul?

By "soul," Cutter does not mean merely an entity that possesses consciousness. For him, the soul is an immaterial entity that is the fundamental bearer or possessor of certain mental capacities. Accordingly, the central question of his article is not whether AI can become conscious, but rather: Could an artificial being possess the same kind of immaterial existence that dualists attribute to human beings?

Cutter's other assumption is that, in the future, a form of AGI will be created that matches or even surpasses human beings in all or nearly all cognitive domains in terms of behavior and function. He emphasizes that this definition of AGI is purely behavioral and functional. Thus, phenomenal consciousness or an authentic and genuine mental life is not part of the concept of AGI as he defines it. For the purposes of his argument, he assumes that it is sufficient to argue for the ensoulment of a minimal form of AGI. His assumption is that such machines could perform virtually any behavioral task as well as, or better than, human beings. He also assumes that the internal functional organization of these systems will in some way resemble that of the human brain. This resemblance need not, of course, be physical or biological; it may exist merely at the level of organization and functional relations (Cutter 2025, 1). On this basis, it appears that Cutter's discussion of AGI ensoulment is grounded in an approach more closely resembling the Emulating the Human Mind approach described above.

On this basis, the central issue in Cutter's article is not the possibility of constructing AGI, but rather its metaphysical status. If AGI is eventually realized, and if human beings genuinely possess souls, should we not entertain the possibility that these artificial systems might also possess souls? Cutter's answer is affirmative. He does not claim that AGI becoming ensouled is the most probable scenario, but he maintains that its probability is sufficiently high that it cannot simply be ignored. His assignment of a 25 percent probability to the hypothesis indicates that he does not intend to defend his claim dogmatically; rather, he seeks to show that the AI Ensoulment Hypothesis should not be classified among highly improbable hypotheses (Cutter 2025, 1–2).

Cutter seeks to show that the two principal assumptions underlying the article – namely, that substance dualism is true and that human beings will ultimately succeed in constructing AGI – are sufficiently reasonable to make discussion of "AI ensoulment" worthwhile. If either of these assumptions were clearly false, the question of AI ensoulment would likewise lose its relevance. In his view, however, neither assumption can be easily dismissed (Cutter 2025, 2).

It is important to note that Cutter deliberately defines AGI in behavioral

and functional terms. Here, "behavior" and "function" refer primarily to cognitive behavior and function, such as problem-solving, learning, reasoning, planning, language use, the performance of cognitive tasks, and the like, but not necessarily to experiences such as pain, color experience, phenomenal consciousness, or first-person experience. Thus, Cutter seeks to ask: Can the mere presence of advanced cognitive functions provide a basis for ensoulment, even if we are still uncertain whether the entity possesses phenomenal consciousness or is a subject?

In defending the assumption of substance dualism, Cutter appeals to the well-known problem of consciousness. In his view, it remains unclear how consciousness can be explained solely in terms of physical processes. This difficulty may provide motivation for accepting some form of dualism. He then presents several considerations that, in his view, make substance dualism superior to property dualism, including its purportedly better account of personal identity, the unity of consciousness, and certain other metaphysical problems (Cutter 2025, 2). With respect to AGI, Cutter likewise argues that the pace of progress in AI technology has been sufficiently rapid that the possibility of achieving artificial general intelligence in the coming decades cannot reasonably be dismissed (Cutter 2025, 2–3).

Cutter rightly points out that taking seriously the possibility that the AI Ensoulment Hypothesis might actually be realized is important and has significant implications. He moves from a metaphysical discussion of the soul to its ethical and practical consequences; in fact, he seeks to show that even if the probability that AI will be ensouled is only "moderate," it nevertheless cannot be ignored, because its ethical implications are far-reaching. Cutter emphasizes that his claim is highly cautious. He does not argue that AI will certainly become ensouled, or even that this outcome is highly probable. His only claim is that we should not regard the possibility as negligible. Yet even this modest claim is of considerable ethical significance.

If we assign some probability to the possibility that AI may possess a soul and, consequently, some form of intrinsic value or moral status, then we cannot treat it merely as a tool. From this perspective, a strong Moral Risk Argument is at work: if AI is ensouled, it is presumably endowed with significant moral status. We cannot therefore regard the probability of its being ensouled as negligible. Ignoring the interests of an entity that may possess moral status could constitute a serious moral error. Accordingly, we should assign some weight to the possible interests of AI in our decision-making. Under such circumstances, its possible interests, suffering, or preferences should also be incorporated into our moral calculations.

Nevertheless, Cutter emphasizes that the current evidence is insufficient to

warrant a high degree of confidence in the AI Ensoulment Hypothesis. In his view, just as decisively rejecting the hypothesis is unwarranted, decisively accepting it is unwarranted as well. The appropriate position, therefore, is for the time being a form of suspension of judgment, accompanied by taking the possibility seriously and assigning it a non-negligible probability (Cutter 2025, 3).

What Is Meant by Soul?

Before discussing "AI ensoulment," Cutter clarifies what he means by "soul" and what he does not presuppose by this concept. Here, Cutter seeks to provide a minimal and relatively inclusive definition of the soul, one that can be shared by different formulations of dualism. He regards the soul as an immaterial entity that in some way underlies particular mental capacities and faculties, such as consciousness, thought, and rational agency, while at the same time being united with a physical system.

The immateriality of the soul, in general, means that the soul is not made of anything physical; it is not composed of atoms, molecules, or anything else investigated in microphysics. Cutter merely takes the soul to be the basis of mental capacities and deliberately refrains from entering into certain deep metaphysical disputes. He does not intend to address the question of clarifying the precise relation between the soul and mental properties. For example, he does not clarify whether the soul itself is the subject that thinks and chooses, or whether the actual subject is something else – perhaps an entity composed of body and soul – to which agency can nevertheless be attributed by virtue of the soul's functioning, just as it is the person who breathes through the activity of their lungs, and it would be incorrect to say that the lungs themselves breathe.

Nor does he take a position on the question of whether the person is identical with the soul or is a more encompassing entity of which the soul constitutes only one part. This question is, of course, highly important and lies at the heart of the disagreement between the Cartesian and Aristotelian views of the soul. Cutter likewise does not address several other disputed issues, such as whether the soul is responsible for non-mental vital functions, such as nutrition, growth, and digestion, or whether non-human animals also possess souls.

In this way, he seeks to provide a definition of the soul that is acceptable both to Cartesian dualists and to proponents of certain other theories of the soul. He states that, for his present purposes, it is sufficient to take substance dualism to entail that the soul is a nonphysical substance, that is, a particular concrete object distinct from every wholly physical object, event, or process (Cutter 2025, 3–4).

What Is Meant by Ensoulment?

Cutter defines the concept of ensoulment operationally. In effect, he seeks to determine precisely when a physical system can be said to "possess a soul." In his view, the mere existence of a soul somewhere in the universe is not sufficient; rather, there must be a direct and stable causal connection between the soul and a physical system. He conceives of this connection as operating in both directions. On the one hand, bodily states can affect states of the soul; for example, neural processes may give rise to sensory experiences (upward causation, or body-to-soul causation, bottom-up). On the other hand, the soul can also affect the body; for example, an individual's volition may cause bodily limbs to move (downward causation, or soul-to-body causation, top-down). Any system that stands in such a regular causal relationship with a soul counts, in Cutter's view, as "ensouled":

> a physical system (like a physical computer) is "ensouled" (or "has a soul," or is "united to a soul") only if there are direct causal connections of this kind between the physical system and a soul, such that the states or activities of the physical system directly and systematically affect the soul, or vice versa (inclusive "or") (Cutter 2025, 4).

Accordingly, the central question becomes: Can an AGI, in addition to being a physical system, possess a soul such that this kind of causal interaction obtains between the two?

Conditions for Ensoulment

After specifying what he means by ensoulment, Cutter argues that ensoulment is neither arbitrary nor accidental. If substance dualism is true, we should expect there to be a lawful, reliable, and counterfactual-supporting system such that, if conditions C obtain in a physical system, a soul will be united with it (or the system will be ensouled). Conditions C might consist of a particular set of states and properties of a human embryo (or, in accordance with what ancient Peripatetic philosophers held, the formation of a particular human temperament). Cutter's important point is that conditions C must be conditions that are sufficient for ensoulment according to the laws of nature – that is, a nomologically sufficient condition for ensoulment.

On this basis, the Ensoulment Hypothesis would be roughly equivalent to the following hypothesis: the matter constituting some future AI systems possesses conditions that are nomologically sufficient for ensoulment. Now, if we are speaking about the ensoulment of an intelligent system that is functionally similar to a human being, conditions C would constitute functional

sufficiency; that is, possessing a human-like functional structure and organization would be a condition that is nomologically sufficient for ensoulment. Cutter maintains that the two arguments he presents in support of the Ensoulment Hypothesis lend support to something resembling the Functional Sufficiency Hypothesis. Cutter identifies various possible structural and organizational similarities between a machine and a human being that could be sufficient for ensoulment, without attempting to determine precisely what those necessary similarities are (Cutter 2025, 5).

Thus, in Cutter's view, human beings are ensouled because the human body possesses certain distinctive natural conditions. He takes these conditions to be of the kind constituted by functional organization, because if human ensoulment were contingent upon the particular material constituting the human body, then talking about the ensoulment of an intelligent machine made of different materials would be meaningless and pointless. Thus, there is something in the human body, and something that has occurred in it, that accounts for human ensoulment. That something is not the material constituting the human body itself, but rather its functional organization and structure. If, then, a machine is to be ensouled, it must resemble a human being in this respect. The reasoning implicit in Cutter's account can be formulated as follows:

1. There is a law-governed condition (C) for ensoulment.
2. That condition is possession of a particular (human-like) functional organization.
3. Assumption: AGI will possess that functional organization.
4. Conclusion: AGI will satisfy the sufficient condition for ensoulment and will therefore be ensouled.

One critical question concerning Cutter's view and the path he takes is why functional organization should be sufficient for ensoulment. Do distinctive human biological structures or particular chemical compositions of the human body not play any role in ensoulment? Cutter appears to have adopted a kind of functionalist presupposition concerning cognition and does not regard ensoulment as dependent on specifically human biological matter or structure, as noted earlier. Whether this presupposition is correct or mistaken remains an open question.

Cutter is not, of course, a functionalist in the sense that he identifies the soul with a function. In his view, which connects ensoulment with substance dualism, functional organization is a condition for becoming ensouled. Thus, on his account, the soul is something distinct from a function or from functional organization.

Cutter makes an important point concerning the relation between the

nomologically sufficient conditions for ensoulment and the actual coming into existence of a soul. The kind of lawfulness Cutter has in mind is nomological lawfulness, which does not necessarily imply natural laws. According to the anti-emergentist view, natural entities can never be the cause of the coming into existence of a soul; rather, they merely provide the occasion or conditions for God to create a soul and infuse or unite it with a natural entity. On this view, such nomological regularity does not express a natural law, because if it were a law of nature, the soul would have to be an effect of the physical natural entity itself, as emergentists maintain.

Nevertheless, even on an anti-emergentist view, it is possible to speak of conditions that are nomologically sufficient for ensoulment – conditions that would constitute the occasioning cause for the divine creation of a soul. With these considerations in mind, Cutter emphasizes that the AI Ensoulment Hypothesis does not entail that we human beings, who are the creators of intelligent systems, create the souls of future AGIs. If future machines are ensouled, our role is merely to provide the material conditions for ensoulment. Whether we are the creators of those souls is an open question; it depends on our metaphysical view concerning the true cause of the soul.

For Cutter, what matters is that, once conditions C are realized, a soul comes to be realized, whether through natural emergence or through special divine creation. In the case of an ensouled human being, whenever certain physical and material conditions are realized – when a human embryo is formed and develops under appropriate conditions, whether in the womb through natural sexual reproduction or in a laboratory through other methods of combining gametes – the human body becomes ensouled. This process is law-governed and supports a counterfactual conditional: if such conditions were not realized, ensoulment would not occur.

The same holds in the case of the ensoulment of a machine: such an order and regularity obtain, and ensoulment is not a random or arbitrary event. Moreover, this account does not depend on any particular metaphysical view concerning the coming into existence of the soul or its relation to the body. For example, it is compatible both with theories according to which the soul comes into existence prior to the body and with theories according to which the soul comes into existence together with the body. One might hold that souls already exist and that material conditions merely provide the occasion for a soul to become united with a particular body, rather than providing the conditions for the creation of a new soul.

Cutter maintains that whatever is true of human beings must also be true of machines. Programmers and engineers provide certain physical conditions, and the artificial intelligence then becomes ensouled, whether the soul is newly

generated and the physical conditions are the cause or occasioning cause of its coming into existence, or whether the soul already exists and the physical conditions are the cause or occasioning cause of its becoming attached or united to the machine's physical system (Cutter 2025, 5–6).

The conclusion, therefore, is that ensoulment (being united with an immaterial soul) is ontologically distinct from any functional property of a physical system (i.e., the Ensoulment Hypothesis is entirely different from functionalism), although certain functional properties may be nomologically sufficient for ensoulment (Cutter 2025, 7).

The Fitting-Recipient Argument

Cutter offers two arguments in defense of the AI Ensoulment Hypothesis, one of which is the "Fitting-Recipient Argument." This argument is based on the idea that, given the assumptions discussed above, whenever a physical system is "fit" to possess a soul, it will become united with a soul (i.e., it will become ensouled). Roughly speaking, "being fit to possess a soul" means that the physical structure of a system is such that it can enter into meaningful interaction and cooperation with the activities of a soul.

Cutter formulates the argument as follows:

Premise (1). Being fit to receive a soul is, in accordance with the laws of nature, a sufficient condition for a physical system to become united with a soul.

Premise (2). A machine that is functionally similar to a human being (i.e., a machine that is appropriately similar to the human body/brain in terms of its functional organization and behavioral capacities) is fit to possess a soul.

Conclusion. A machine that is functionally similar to a human being will become united with a soul (Cutter, 2025, pp. 18–19).

Before defending the two premises of the argument, it is necessary to say something about the meaning and criterion of a body's fitness to receive a soul. Whether a body is fit and suitably constituted to receive a soul depends, to some extent, on the nature and characteristics of the body and, to some extent, on the nature and characteristics of the soul. More specifically, this fitness is a function of the soul's fundamental or basic powers and activities. Cutter seeks to argue that if a machine possesses powers similar to those of the human body, it will likewise be fit to receive a soul.

The human soul has certain distinctive basic powers or capacities, including

sensory experience, judgment, practical and theoretical reason, and choice. To better understand when a physical system is suitable for such a soul, we should consider cases in which the states and capacities of the human body complement or support the fundamental powers of the soul. In other words, to determine what makes a body a suitable recipient of a soul, we must first determine what the soul does and then examine what kind of body could cooperate with and be suited to the performance of those activities. If the soul possesses sensory perception, thought, decision-making, and choice, what kind of physical system could effectively interact with those powers? (Cutter, 2025, p. 19).

For example, if we want to determine what kind of vehicle is suitable for a pilot, we must first identify the pilot's basic activities (e.g., navigating and controlling a physical object in the air), and we can then determine what kind of structure an aircraft must have in order to make these activities possible. Cutter appears to apply the same model to the soul. He offers an argument along roughly the following lines:

1. The soul possesses cognitive powers;
2. The human body supports these powers;
3. AGI can likewise provide the same kind of support.

Conclusion: AGI can likewise be a suitable recipient for a soul.

Cutter makes several points in an effort to show why the human body is a suitable substrate for the human soul. First, the human body provides the inputs necessary for sensory consciousness. The human body possesses a remarkably rich set of internal states, particularly the neural representations present in the brain's perceptual systems, that reliably correlate with states of the external environment. These physical states can serve as the direct causal basis of the soul's sensory experiences. As a result, through these states, the soul can acquire a rich and accurate representation of its surroundings and, on that basis, form a large number of beliefs about the world that are generally accurate and rational. Most other physical objects, such as rocks or chairs, lack such a structure. A rock does not contain a set of internal states capable of providing appropriate inputs for sensory experience through which a soul could construct an accurate and reliable model of its environment (Cutter, 2025, p. 19). Cutter's argument can be summarized in a single sentence: the human body is useful to the soul because it provides the soul with rich and structured information about the world; the human body is not merely a container for the soul, but a complex information-processing system that continuously supplies the soul with valuable data.

Second, the human body provides its soul with a remarkably broad and

diverse range of behaviors through which the subtleties and complexities of its mental life can be expressed. Since human outward behavior is ultimately produced through the brain, the capacity to express such mental subtleties requires the brain to possess a structure capable of reflecting or encoding them (Cutter, 2025, p. 19). In the previous example, the human body was presented as a source of input for the soul; here, it is presented as a means of obtaining output from the soul. On the one hand, through the body, the soul receives information from and is causally affected by the world; on the other hand, through its body, the soul expresses itself in the world and exerts an influence upon it.

Cutter cites a similar point from J. P. Moreland in explaining why bodies are created. Moreland argues that the more complex an entity's consciousness is, the more complex and sophisticated its body must be in order to enter into causal interaction with the subtle and diverse states of that entity's mind. If an entity possesses highly complex consciousness but is connected to a very simple material object, much of the complexity of its mental life would never have the opportunity to manifest itself or be realized outwardly and would therefore remain without practical effect. For this reason, objects such as rocks or chairs are not suitable recipients for souls. Even if a soul with a complex inner life were attached to a rock, the rock's structure would not allow that complexity to be manifested in its behavior. A rock possesses neither a sufficient diversity of behavior nor the internal organization required for the subtle regulation of behavior.

Third, the human body assists the soul in reasoning and thinking. The brain is organized in such a way that it may play a supportive role in the exercise of the soul's rational faculties. It provides a rich store of memories encoded in neural networks that the soul can draw upon. The brain also contains temporary representations held in working memory, which can facilitate reasoning and function as a kind of "neural scratchpad" or "neural notepad."

A dualist would presumably maintain that no sequence of physical events in the brain is itself the process of reasoning. Nevertheless, some of these events can assist or complement reasoning, just as notes written on a sheet of paper are not themselves part of the reasoning process but can help us carry out that process. By contrast, objects such as rocks or chairs typically do not possess internal states capable of playing such a supportive role in reasoning. Thus, compared with the human brain, a rock is far less well suited to cooperate with the rational faculties of a soul (Cutter, 2025, p. 20).

Defense of Premise 1

After presenting examples of the cooperation between body and soul, Cutter

seeks to extract a general principle from them, namely, to defend the reasonableness of the first premise of the argument (being fit to receive a soul is a sufficient condition for becoming ensouled). The human body is organized in various ways that enable it to facilitate and cooperate with the activities of the soul. It can therefore be said that the human body is "fit to receive a soul."

Moreover, it seems that the natural explanation of the fact that a living human body possesses a soul, whereas rocks and human corpses do not, lies precisely in this difference in their fitness and capacity to cooperate with the activities of the soul. The living human body possesses a structure that enables it to cooperate effectively with the activities of the soul, whereas rocks and corpses lack such a capacity. Cutter then advances a stronger claim: he suggests that even our intuitions about actual and hypothetical cases of ensoulment are implicitly based on this very criterion. For example, if we accept dualism concerning human beings, we often find it natural to suppose that intelligent extraterrestrial beings would also possess souls. The likely reason for this intuition is that we imagine their bodies to be organized in such a way that they can cooperate with the activities of a soul. By contrast, although it is logically possible to suppose that a rock, a chair, or a corpse possesses a soul, we do not take such suppositions seriously because these objects lack the physical organization necessary for meaningful cooperation with the activities of a soul. Taken together, these considerations support the idea that "fitness to receive a soul" is a sufficient and law-governed condition for becoming ensouled (Cutter, 2025, pp. 20–21).

Cutter goes on to explain that the fitness and suitability described above are gradable, admitting of degrees and variations in strength and extent. A physical system can possess a structure that enables it to cooperate with the activities of a soul to different degrees; this is true even among human beings. For example, brain damage or developmental impairments reduce the degree to which a body is suited and fit to cooperate with the activities of the soul (Cutter, 2025, p. 21).

Cutter is addressing here an objection that can readily be raised: if "fitness for a soul" is the criterion for ensoulment, precisely at what point does this fitness begin? Fitness to receive a soul appears to be a gradable property. A physical system can possess, to varying degrees, a structure that enables it to cooperate with the activities of a soul. This raises the question of where the boundary or threshold required for ensoulment should be drawn.

Cutter's response is that, for present purposes, the threshold can be set very high, namely, at the level of an ordinary, mature human body. Since the first premise of the argument provides only a sufficient condition for ensoulment, setting such a high threshold does not entail that creatures below this level lack souls. It merely claims that any system that reaches this degree of fitness will

thereby satisfy the conditions for ensoulment (Cutter, 2025, p. 21).

In fact, Cutter recognizes that the concept of a "fitting recipient" is vague and admits of degrees, because fitness is not a binary property; it is not, for example, like being pregnant, which either obtains or does not. Rather, it is something more like being intelligent or being healthy. If fitness is gradable and admits of degrees, however, what about a chimpanzee? A dolphin? A human infant? A human fetus? A person with severe brain damage? A partially advanced AGI? At what point does ensoulment occur?

Cutter does not answer these questions. Instead, he argues that, for present purposes, we do not need to know precisely where the boundary of ensoulment lies. It is sufficient to say that an ordinary mature human being has clearly crossed the threshold. A human-like AGI will cross it as well. Thus, discussion of borderline cases can be set aside for the time being, as Cutter seeks to focus only on clear cases. He does not claim that only beings with a high degree of fitness possess souls; rather, he claims that any being that reaches this level of fitness will possess a soul. For his argument, it is sufficient to accept that the human level of fitness is sufficient for ensoulment.

Defense of Premise 2

Thus far, Cutter's efforts have been directed toward showing that fitness is a sufficient condition for ensoulment and that the human body is suitable and fit for ensoulment. He now seeks to take the second important step and argue that what gives rise to this fitness is nothing more than functional organization. The second premise of the argument is that a functionally human-like machine will be suitable for receiving a soul. The reason for this claim is that fitness to receive a soul depends entirely on the functional organization and behavioral capacities of a system. All the examples previously given to explain the suitability of the human body for the soul concerned only the functional and behavioral features of the human body, not the particular material of which the body is composed. For example, it was argued that the brain possesses a rich set of internal states that can serve as inputs for sensory experiences and make possible the formation of a rich model of the environment. But this feature is not dependent on the biological nature of the brain. A computer could likewise possess similar internal states, even if those states were realized in a non-biological substrate. The same point applies to the other examples as well. Thus, if fitness to receive a soul depends solely on functional organization and behavioral capacities, and if the human body is suitable for the soul because of precisely these features, then a machine that is functionally similar to a human being will likewise be equally suitable for receiving a soul (Cutter, 2025, p. 21).

The most important concept here is substrate independence, on which the entire argument is effectively based. The idea is this: suppose that the human brain is composed of neurons. What is relevant to and plays a role in ensoulment? The first option is the biological material of the neurons; the second is the pattern of relations and functions realized by the neurons. Cutter chooses the second option. He argues that if something can perform the same roles as neurons – that is, process the same information, possess the same representations, and produce the same behaviors – it will possess the same fitness for ensoulment. Thus, for present purposes, it makes no difference whether the physical system is made of neurons, silicon chips, or some other material.

Cutter refers precisely to the examples he himself presented earlier. In any of those previous examples, did he refer at any point to carbon, DNA, or proteins? The answer is no. All of the examples concerned information, representation, memory, behavior, reasoning, and the like. Why, then, should biology suddenly become relevant?

Critical Evaluation

Cutter employs the approach of substance dualists to establish the ensoulment of intelligent machines. Until now, substance dualists have established ensoulment only in the case of natural (i.e., non-artificial) beings, and it may previously have been assumed that an artificial being could not become ensouled, or that there was no reason to regard its ensoulment as even possible. In the case of a natural human being, the substance dualist maintains that whenever an appropriately constituted human body comes into existence, it will become ensouled, according to various metaphysical explanations that have been offered in this regard, some of which were mentioned above.

The particular metaphysical explanation of how a human body becomes ensouled is independent of the conclusion Cutter seeks to establish. His claim is that, for the same reason and through the same process by which an appropriately constituted human body becomes ensouled, an intelligent machine will also become ensouled, and that there is no relevant difference between the two cases. In other words, if substance dualism is true of human beings, it should also be true of human-like machines, because what matters is the presence of certain distinctive mental powers and states together with a body suitably constituted for those powers and states – and both conditions obtain in the case of both natural human beings and AGI.

Accordingly, we should posit a general principle common to human beings and AGI. If we call the concept common to the human soul and the AGI soul

soul*, and the concept common to the suitably constituted human body and the suitably constituted physical body of a machine (which are functionally similar to one another) body*, then:

Whenever body* is realized, it becomes united with soul*.

Anyone who rejects this general principle and maintains that a machine cannot become ensouled, thereby objecting to the Fitting-Recipient Argument, must, if they find the first premise acceptable on a dualist approach, regard the functional similarity between the machine and the human body as insufficient for the machine's ensoulment – that is, they must reject the second premise. In that case, they must introduce some additional condition or conditions for ensoulment that are realized in human beings but cannot be realized in machines, and consequently conclude that a machine will never become ensouled.

Békefi (2026) adopts this strategy in criticizing Cutter. He argues that some additional condition must be involved and calls this condition "Integrity." Accordingly, Békefi's objection to Cutter is what he calls "The Missing-Integrity Objection.[1]" Békefi accepts the first premise of the Fitting-Recipient Argument and directs his objection against its second premise. However, Békefi's strategy does not seem adequate for two reasons. First, he does not address the central problem with the argument, namely, treating a necessary condition as a sufficient one. Second, his criticism is open to response, as he himself acknowledges at the end of his critique of the Fitting-Recipient Argument:

If engineers could build a neural network using billions of memristors of that kind, and manage to create for it an environment suitable for the kind of deep learning characteristic of both animals and artificial neural networks, it is conceivable that a system could be produced with both human-level intelligence and high integrity. It would not be easy to decouple its physical parts and functional capacities (though the danger of being turned off would likely remain). I have no objection to the ensoulment of such an entity – but I suspect that rather than falling under our ordinary concepts of 'AI' and 'machine' (cf. Section 1.3), it would better be classified as some other thing – perhaps a synthetic brain. (Békefi, 2026, p. 134)

Rejection of Premise 1

As discussed above, "being fit to possess a soul" roughly means that the

1. I leave the examination of this objection for another occasion.

physical structure of a system is such that it can enter into meaningful interaction and cooperation with the activities of a soul. According to the first premise of the Fitting-Recipient Argument, if fitness obtains in a physical system, that system will become ensouled without exception or violation.

In explaining the truth and reasonableness of this premise, Cutter argues that the human body is fit to receive a soul and that the natural explanation of the fact that the living human body (and every other living body) possesses a soul, whereas certain objects, such as rocks and chairs, do not, lies in their difference in fitness for ensoulment. In other words, we have encountered ensouled entities, and we have explained their conditions and their difference from soulless objects by holding that they are ensouled because their bodies are fit to receive a soul. Fitness is a characteristic of an ensouled body precisely insofar as it is ensouled, and it expresses a particular relation between a body and its soul. This means that bodily fitness plays a necessary and effective role in ensoulment, in the sense that, had such fitness been absent, ensoulment would not have occurred. This state of affairs can be expressed by the following conditional proposition:

(I) If an entity is ensouled, then it has a body that is fit to receive a soul.

Proposition (I) means that ensoulment entails having a suitably constituted body; in other words, having a body that is fit to receive a soul is a necessary condition for ensoulment, while ensoulment is, conversely, a sufficient condition for having a body that is fit to receive a soul. (Here, necessary and sufficient conditions are used in accordance with standard philosophical theory; see Brennan, 2024).

By contrast, the first premise of the Fitting-Recipient Argument is:

(II) If an entity has a body that is fit to receive a soul, then it is ensouled.

Proposition (II) means that having a body that is fit to receive a soul is a sufficient condition for ensoulment, while ensoulment is a necessary condition for having a body that is fit to receive a soul.

The crucial issue that undermines the argument is that while proposition (II) serves as the actual premise of the argument, Cutter's reasoning can at most establish proposition (I), which is the converse of (II). In other words, what is true and what can be supported by argument is proposition (I) – that having a body fit to receive a soul is a necessary condition for ensoulment – yet the converse of this proposition has been introduced as a premise of the

argument, namely, that having a body fit to receive a soul is a sufficient condition for ensoulment. The distinction between the two has been conflated, resulting in a logical fallacy. This fallacy is known variously as affirming the consequent, the converse error, or treating a necessary condition as sufficient.

In fact, the argument would have to be reformulated as follows:

(I) If an entity is ensouled, then it has a body that is fit to receive a soul.

Premise (2). AGI has a body that is fit to receive a soul.

Conclusion. AGI is ensouled.

Formulated in this way, it is obvious that this argument commits the formal fallacy of affirming the consequent.

Indeed, one may accept that fitness for ensoulment is necessary for ensoulment, and that if such fitness is absent, ensoulment will not occur. However, it does not follow that fitness is all that is required for ensoulment. Fitness is a necessary condition for ensoulment, but it is not a sufficient condition. Therefore, even when fitness is present, one cannot infer that ensoulment will occur.

Thus, although AGI may be fit for ensoulment by virtue of its functional similarity to the human body, it does not follow from this fitness that AGI will necessarily become ensouled. Accordingly, the Fitting-Recipient Argument for the ensoulment of artificial intelligence fails.

Therefore, although the Fitting-Recipient Argument, as formulated by Cutter, is formally valid, its first premise appears to be problematic. According to the first premise, having a body that is fit to receive a soul is a sufficient condition for ensoulment. However, the argument and explanation Cutter offers in support of it can establish only, at most, that having a body that is fit to receive a soul is a necessary condition for ensoulment. Now, if we revise the first premise by replacing the sufficiency of having a fitting body for ensoulment with its necessity, the argument becomes formally invalid, since it would then commit the fallacy of affirming the consequent and would no longer establish its conclusion.

It might be argued that Cutter, as he himself claims, has successfully argued and explained that having a body fit to receive a soul is a sufficient condition for ensoulment, and that the first premise is therefore true. However, this does not seem to be the case. To clarify why Cutter's explanation and argument establish, at most, that fittingness is a necessary condition for ensoulment, let us examine the possible relations between fittingness and ensoulment in the

case of human beings:

Fittingness for ensoulment is either the sole condition for the occurrence of ensoulment (in which case it would be both a necessary and a sufficient condition) or one of several conditions required for ensoulment to occur. Cutter does not claim, nor do his explanation and argument show, that fittingness is the sole condition for the occurrence of ensoulment. Moreover, this assumption appears to be incompatible with the metaphysical commitments and understanding of the causal relation between soul and body held by some substance dualists (recall Cutter's point concerning emergentism and anti-emergentism: if ensoulment were made to depend wholly and lawfully on the instantiation of a particular bodily state, we would be moving toward an emergentist theory of the soul–body relation. Of course, some philosophers endorse emergentist dualism; see Hasker, 2018; for a critical discussion, see Rickabaugh, 2018. For a contemporary formulation and defense of Cartesian substance dualism, see Swinburne, 2018). Thus, assuming that fittingness is only one of the conditions required for ensoulment to occur – and Cutter agrees with this assumption and does not link the truth of the AI Ensoulment Hypothesis to the uniqueness of the fittingness condition – we shall proceed with the discussion.

If the occurrence of ensoulment requires the instantiation of $n$conditions and fittingness is one of them, it does not follow merely from this fact that fittingness is a sufficient condition for ensoulment. For fittingness to be a sufficient condition, it must be shown that whenever fittingness obtains, the other conditions necessary for ensoulment will necessarily obtain as well. Only then can one infer from the instantiation of fittingness to the instantiation of the full set of conditions necessary for ensoulment and, consequently, to the occurrence of ensoulment. In such a case, it may, with some qualification, be said that fittingness is a sufficient condition for the occurrence of a soul in a physical body, since the other necessary conditions are guaranteed by its instantiation. But is the assumption that the instantiation of fittingness also guarantees the instantiation of the other conditions necessary for ensoulment tenable? Cutter's argument and explanation do not suffice to establish this claim. At most, Cutter shows that human beings are ensouled and that, in cases where ensoulment occurs, the human body is fit to receive a soul. He says that the natural explanation of the fact that a living human body has a soul, whereas rocks and human corpses lack souls, appears to be precisely this difference in their fitness and cooperation with the activities of the soul. But what does he mean by "natural explanation"?

Cutter may mean by "natural explanation" a form of inference to the best explanation. In that case, in explaining the relation between fittingness and

ensoulment in human beings and other ensouled entities, as well as the relation between unfittingness and the absence of ensoulment in entities such as rocks and chairs, he appeals to a form of inference to the best explanation and takes fittingness to be the best explanation of this pattern of ensoulment and non-ensoulment. However, the fact that fittingness is the best explanation of this pattern does not, in itself, entail that fittingness is either a sufficient or even a necessary condition for ensoulment.

Suppose that all existing instances of Ahave two properties, $a$and b, and that $a$enjoys a certain priority over b(more precisely, that A's having $a$is causally prior to A's having b), such that the best explanation of why instances of Ahave property $b$is that they have property a. It does not follow from this explanation that the presence of $a$is a sufficient condition for the presence of b. To establish that $a$is a sufficient condition for b, it must be shown that necessarily (or in every possible world), anything that has $a$also has, or must have, b. Mere observation of the co-occurrence of $a$and $b$is not sufficient here; rather, it must be shown why the instantiation of $a$entails the instantiation of b. In other words, it must be shown that the non-instantiation of b is incompatible with the instantiation of a – that is, that it is impossible for an entity to have $a$without having b(in the present case, that it is impossible for an entity to have a body fit to receive a soul without being ensouled).

Cutter, of course, claims that he is going to establish precisely such a claim; that is, he intends to establish that there is a natural nomological relation between fittingness and ensoulment. The point, however, is that his argument for the first premise cannot accomplish this. Inference to the best explanation does not yield such a conclusion. In fact, generalizing from actual instances to metaphysical sufficiency is problematic unless it is shown that there is an essential or necessary relation between fittingness and ensoulment. It is not impossible for a body fit to receive a soul to exist without becoming ensouled, since ensoulment may be conditional upon some further condition or conditions, at least one of which may fail to obtain.

If fittingness, C, is a sufficient condition for ensoulment, S, then:

$$C \rightarrow S$$

Now, if ensoulment requires $n$conditions:

$$C_1 \wedge C_2 \wedge \cdots \wedge C_n \rightarrow S$$

and fittingness is one of them – for example:

$$C_1 = C$$

then, for ensoulment to occur, it must be the case that whenever C obtains, $C_2, \ldots, C_n$ also obtain. Accordingly, we must have:

$$C \rightarrow (C_2 \wedge \cdots \wedge C_n)$$

That is, fittingness must guarantee the instantiation of the other conditions necessary for ensoulment as well. But Cutter's argument cannot establish this. Unless this relation is established, C is, at most, a necessary condition rather than a sufficient one.

In fact, Cutter appears to be confronted with a pattern of explained correlation or co-occurrence: every instance of ensoulment is accompanied by fittingness. He takes fittingness to explain ensoulment. However, for fittingness to be a sufficient condition for ensoulment, he would have to show that no other independent metaphysical condition is required for ensoulment. Cutter's argument does not establish this, and in fact, establishing it would not be easy, if it is possible at all. To move from an inferential explanation to such a metaphysical claim, further metaphysical premises or principles would have to be provided that establish a necessary relation between fittingness and ensoulment.

However, Cutter may mean by "natural explanation" a natural nomological explanation, as discussed above. From Cutter's perspective, the AI Ensoulment Hypothesis is roughly equivalent to the hypothesis that the material constituting an ensouled physical entity – whether a natural or an artificial entity – possesses conditions that are sufficient for ensoulment in accordance with the laws of nature, and he identifies these conditions with "functional sufficiency." But in that case, the argument appears to beg the question, since the Fitting-Recipient Argument is supposed to support the Functional Sufficiency Hypothesis, while, in justifying its first premise, it appeals to a "natural explanation" – that is, precisely the natural nomological relation and functional sufficiency that it was supposed to establish.

Conclusion

Cutter's account and argument for the ensoulment of artificial intelligence are valuable and noteworthy. As Békefi puts it, "Cutter's is the only detailed discussion of AI ensoulment in recent memory; indeed, perhaps the first serious treatment of the question in analytic philosophy since Turing's half-hearted comments" (Békefi, 2026, p. 121).

Cutter has sought to explain the AI Ensoulment Hypothesis clearly by articulating its underlying assumptions and key concepts, and he has also done a commendable job of formulating the premises of the Fitting-Recipient Argument. He has constructed an argument for the ensoulment of machines in the form of a *modus ponens*. However, the argument appears to face a significant problem. In effect, he commits the fallacy of affirming the consequent: the first premise of the argument, which he treats as a sufficient condition, is in fact established only as a necessary condition, and his argument therefore commits the fallacy of treating a necessary condition as sufficient.

In fact, everything Cutter has offered concerning the relationship between ensoulment and bodily fitness establishes only that fitness is a necessary condition for ensoulment. For his argument to be valid, however, he needs fitness to be a sufficient condition for ensoulment, as he himself explicitly maintains. Yet he does not appear to provide an argument in support of its sufficiency. Consequently, the Fitting-Recipient Argument does not succeed in establishing the ensoulment of AI.

▣ Conflict of Interest

The author declares no competing interests.

References

Békefi, B. (2026). No ghost in the machine: Doubting AI ensoulment. *Faith and Philosophy: Journal of the Society of Christian Philosophers*, *42*(1), Article 7, 121-146 https://doi.org/10.37977/faithphil.2026.42.1.7

Brennan, A. (2024). Necessary and sufficient conditions. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford encyclopedia of philosophy* (Summer 2024 ed.). Stanford University. https://plato.stanford.edu/archives/sum2024/entries/necessary-sufficient/

Cutter, B. (2025). The AI ensoulment hypothesis. *Faith and Philosophy: Journal of the Society of Christian Philosophers*, *41*(1), Article 1, 1–26. https://doi.org/10.37977/faithphil.2025.41.1.1

Cycleback, D. (2019). *Philosophy of artificial intelligence*. Bookboon.

Goertzel, B., & Pennachin, C. (Eds.). (2007). *Artificial general intelligence*. Springer Verlag.

Goertzel, B., & Wang, P. (Eds.). (2007). *Advances in artificial general intelligence: Concepts, architectures and algorithms: Proceedings of the AGI Workshop 2006*. IOS Press.

Hasker, W. (2018). The case for emergent dualism. In J. J. Loose, A. J. L. Menuge, & J. P. Moreland (Eds.), *The Blackwell companion to substance dualism* (pp. 61–72). Wiley-Blackwell. https://doi.org/10.1002/9781119468004.ch4

Kristjánsson, K. (2010). *The self and its emotions*. Cambridge University Press.

Leon, F. (with ChatGPT). (2025). *Philosophy of artificial intelligence*. ResearchGate. https://www.researchgate.net/publication/388146273_Philosophy_of_Artificial_Intelligence

McGinn, C. (1999). *The mysterious flame: Conscious minds in a material world*. Basic Books.

Oberg, A. (2023). Souls and selves: Querying an AI self with a view to human selves and consciousness. *Religions*, *14*(1), Article 75. https://doi.org/10.3390/rel14010075

Rickabaugh, B. L. (2018). Against emergent dualism. In J. J. Loose, A. J. L. Menuge, & J. P. Moreland (Eds.), *The Blackwell companion to substance dualism* (pp. 73–86). Wiley-Blackwell.

Swann, J. (2021). Anima ex machina: Can artificial intelligence have soul? In W. H. U. Anderson (Ed.), *Technology and theology* (pp. 201–216). Vernon Press.

Swinburne, R. (2018). Cartesian substance dualism. In J. J. Loose, A. J. L. Menuge, & J. P. Moreland (Eds.), *The Blackwell companion to substance dualism* (pp. 133–152). Wiley-Blackwell.

Tijanovic, T. (in press). The virtual ensoulment hypothesis: Can virtual individuals be ensouled? Dialogo.

Tiwari, S. K., & Tomar, P. (2026). Why AI can never have a soul: A philosophical inquiry into language and consciousness. *Journal of Philosophical Investigations*, *20*(54), 233–252. https://doi.org/10.22034/jpiut.2025.69223.4202

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

From the Rational Requirement and Ontological Reconstitution of Human Agency to More Onerous Duties

Muhammad Heydarpour [1] **Ali Razizadeh** [2]

1. Assistant Professor, Department of Philosophy, Faculty of Literature and Humanities, University of Religions and Denominations, Qom, Iran. **(Corresponding author)** m.heydarpour@urd.ac.ir

2. Assistant Professor, Department of Media Arts, Religion and Media Faculty, IRIB University, Qom, Iran ali.razizadeh@iribu.ac.ir

Abstract

Original Research

This paper aims to show that the rational obligation to employ AI (the first thesis), combined with the transformation of the agent's practical identity (the second thesis), can predictably prepare the ground for more numerous, burdensome, or complex rational or moral duties. This is because an agent's obligations are sensitive not only to the agent's practical identity but also to the agent's epistemic and intentional state. As these dimensions of agency are expanded, the range and stringency of the agent's obligations may likewise increase. Initially, two categories of questions – normative and descriptive – are distinguished from one another. The second part of this paper pursues the thesis that if certain AI systems are trustworthy within specified domains, the human agent is not merely rationally permitted to employ AI, but in a range of significant choices and decisions, rationally required to do so. In the third part, we explain that on the one hand, artificial intelligence is not a merely neutral instrument for antecedently given human ends; on the other, human agency is neither dissolved nor reducible to AI agency. Rather, in long-term human–AI interaction, the conditions under which human beings perceive, deliberate, will, and understand themselves undergo a fundamental reconfiguration. The final section uses an example to show how a more demanding moral obligation can arise from a transformation in the agent's practical identity together with a pre-existing rational requirement.

Keywords:

onerous duties, rational requirement, ontological reconstitution, zetetic norms, practical identity.

Received: 2026/02/26; **Received in revised form:** 2026/05/23; **Accepted:** 2026/07/29; **Published online:** 2026/08/03

Heydarpour, M.; Razizadeh, A. (2026). From the Rational Requirement and Ontological Reconstitution of Human Agency to More Onerous Duties. *Journal of Philosophical Theological Research*, (Special Issue on AI, Consciousness, and Human Agency), 28(3), 233-260. https://doi.org/10.22091/jptr.2026.17879.3731

Copyright © The Author

Introduction

A wide variety of debates and topics has emerged concerning artificial intelligence (AI). Amid these discussions, one can draw a fundamental distinction between two broad categories of problems: normative and descriptive.

The first category encompasses questions such as:

1. How ought AI system designers to engineer these systems such that moral principles are not violated?
2. Can AI agents be regarded as morally responsible agents?
3. What duties do human users have in their employment of AI?
4. Do rational agents, capable of deploying AI, possess stronger reason – on balance against contrary reasons – to incorporate AI outputs as candidate considerations under scrutiny in their deliberation about significant theoretical and practical matters? If so, on what grounds?

The second category raises questions of a different sort:

1. What advances in AI can be expected, and how will they affect economic, cultural, legal, political, and military life?
2. Can AI agents possess all, or the most salient, capacities of human agents?
3. What novel possibilities, hitherto unavailable to human beings, will AI generate?
4. Which ontological conception of human–AI interaction is correct, and why?

In the second part of this paper, we intend to argue for an affirmative answer to the last question of the first category and explain its rationale. This part pursues the thesis that if certain AI systems can merit the trust of a rational agent within specified domains, the human agent is not merely rationally permitted to employ AI, but in a range of significant choices and decisions, rationally required to do so. In the third part, addressing the last question of the second category, we respond that on the one hand, artificial intelligence is not a merely neutral instrument for antecedently given human ends; on the other, human agency is neither dissolved nor reducible to AI agency. Rather, in long-term human–AI interaction, the conditions under which human beings perceive, deliberate, will, and understand themselves undergo a fundamental reconfiguration.

Normative Engagement with AI

To elucidate human normative engagement with AI, it is necessary to examine some of the pivotal mental phenomena an agent undergoes in arriving at choices. Some of what is involved in the process of normative decision-making includes: cognitive states (such as belief, knowledge, and thought), affective states (such as fear, joy, grief, and empathy), conative states (such as intention, desire, and motivation), understanding, reasoning, experience, moral intuition, virtues and vices, character, reflection or deliberation, consciousness, autonomy, and rationality.

A standard line of argument maintains that AI agents lack consciousness, autonomy, rationality, intentionality, a sense of responsibility, and the phenomenological experience of moral obligation. Given that these features are typically conceived as necessary conditions for the attribution of moral responsibility and moral agency, AI systems cannot qualify as moral agents. Conversely, another line of argument – endorsing alternative conceptualizations of the aforementioned criteria and advocating minimalist accounts of agency – seeks to ascribe a species of moral agency to AI systems. In this paper, we aim to show that even if AI agents entirely lack consciousness, autonomy, rationality, intentionality, and the like – and hence cannot be deemed moral agents – competent human agents remain rationally obligated to consult AI systems and critically consider their outputs when making significant normative life decisions.

By "significant life decisions," we refer primarily to accepting fundamental propositions as true and to the deliberation over, and selection of, worthy ends, and secondarily to the determination of means conducive to realizing these ends. By "employment" or "use," we mean the competent deployment of an AI system to generate an appropriate output that functions as a practical consideration to be factored into the process of practical deliberation. Accordingly, the AI's final output need not be treated as an all-things-considered consideration; rather, it constitutes a prima facie (or pro tanto) consideration that requires critical human validation before acquiring decisive, normative authority. (Alternatively, an agent might first arrive at an unassisted all-things-considered judgment through standard human deliberation, and subsequently invoke AI as a critical test – examining the AI's output alongside human reasoning to corroborate, disconfirm, or revise that initial judgment before settling on a definitive, endorsed stance.) Crucially, however, consulting AI systems and treating their output as a prima facie consideration cannot be regarded as an arbitrary matter or a mere subjective preference, for reasons we elaborate below.

To address this question, we may attend to the distinctive comparative advantages AI systems possess over human beings. Among the mental phenomena involved in normative decision-making, AI systems can aid us within the spheres of cognitive states, reasoning, and reflection/deliberation. Specifically, they surpass human agents in high-speed computation, processing massive datasets, categorization, analysis, interpretation, argumentation, inference, problem-solving, and operational stability. Stated differently, humans are subject to numerous cognitive and epistemic limitations that can be mitigated through AI. These limitations include cognitive biases, affective interference (e.g., anger, fear, or partiality), conative distortions, cognitive fatigue, hunger, boredom, oversight, and memory lapses. By employing AI systems, agents can access an expansive repository of human knowledge and experiential records to optimize their choices.

For the sake of clarity and to avoid disputes, we assume that the following claims are true:

1. Rationality is divided into structural rationality (internal coherence among mental states) and substantive rationality (responsiveness to reasons);
2. Rationality is further divided into theoretical rationality and practical rationality;
3. The constitutive aim of theoretical rationality is truth;
4. The constitutive aim of practical rationality is the good;
5. The primary output of theoretical rationality is a cognitive state, such as belief;
6. The primary output of practical rationality is a conative state, such as intention;
7. Theoretical rationality consists – standardly construed – in recognizing and appropriately responding to epistemic reasons;
8. Practical rationality consists – standardly construed – in recognizing and appropriately responding to practical reasons.

1) The Requirement of Theoretical Rationality

Argument 1: Zetetic Norms

Michele Palmira (2024) argues that the rational response to higher-order evidence (HOE) requires that, upon encountering such evidence, we must reopen inquiry (re-investigate) rather than immediately abandoning our current belief. When an agent possesses first-order evidence for a proposition

p, and subsequently receives higher-order evidence that sufficiently casts doubt on the reliability or epistemic credentials of her first-order evidence, she falls under a zetetic obligation to investigate whether $p$is true. This inquiry may be direct – such as re-performing the calculation – or indirect, by investigating whether the higher-order evidence is genuinely reliable. Given the conflict between treating a matter as settled and being under an obligation to re-investigate it, zetetic norms suggest that retaining the belief as settled is no longer justified (Palmira, 2024).

Palmira adduces the following case: Andre calculates the dinner bill in her head and comes to believe that the bill is $45. She then receives highly credible testimony that she may have inadvertently taken a pill that impairs mental arithmetic. This testimony constitutes higher-order evidence: evidence about her own cognitive process and the reliability of her first-order evidence. As it happens, the pill has not actually been effective. This higher-order news does not prove that the $45 verdict is false; nonetheless, it provides Andre with a compelling epistemic reason not to treat the question as settled and to reopen inquiry.

Palmira posits that rather than directly defeating the first-order belief tout court, higher-order evidence can impose a normative obligation to reopen inquiry. Because complying with this zetetic obligation precludes an agent from treating the question as settled, retaining the belief as a closed, definitive verdict becomes rationally impermissible with respect to that obligation. In this manner, Palmira explains the normative significance of higher-order evidence through the lens of norms of inquiry (zetetic norms).

A parallel argument can be reconstructed to establish the rational requirement of employing AI in an agent's significant life decisions:

- **Premise 1 (The Epistemic Principle):** Whenever a rational agent Spossesses first-order evidence for believing that p, but simultaneously has access to strong higher-order evidence indicating a high probability of error or bias in their cognitive faculties, $S$is rationally required to re-examine p.

- **Premise 2 (The Cognitive Condition of Human Agents):** In high-stakes decision-making contexts, any rational human agent S has access to strong higher-order evidence (derived from cognitive science and lived experience) demonstrating that their intuitive and computational processes are constantly susceptible to systematic bias, fatigue, prejudice, and memory blind spots.
- **Premise 3 (From 1 and 2):** In high-stakes decision-making contexts, a rational human agent S has a zetetic obligation to re-examine their

reasoning and decision-making process.

- **Premise 4 (Epistemic Consequentialism):** The obligation to re-examine is a duty that can only be satisfied by employing methods that maximize the chance of detecting existing errors and avoid replicating the initial biases.
- **Premise 5 (The Inadequacy of Exclusive Self-Reliance):** Drawing solely on the agent's own capabilities (conducted without external tools) remains prone to the same prior biases and errors; thus, it fails to optimally satisfy the condition of epistemic efficacy, necessitating the use of external and independent assessment tools or resources.
- **Premise 6 (Partial Cognitive Efficacy of AI):** AI systems, acting as a processing tool independent of the agent's biological psychology, can provide the rational agent S with overlooked options, logical inconsistencies, and counter-evidence at a low cost.
- **Premise 7 (From 4, 5, and 6):** The employment of AI systems is one of the most effective, accessible, and rational means for fulfilling the obligation of re-examination in significant human decisions.
- **Premise 8 (From 3 and 7):** Therefore, rational agents are under a rational obligation to employ AI outputs as an effective tool for re-examination and the critical assessment of evidence in the process of their belief-formation.

Argument 2: Epistemic Blame and Normativity

According to the traditional view in epistemology, epistemic norms strictly govern the response to evidence the agent already possesses; gathering further evidence or conducting research falls solely under the purview of practical, instrumental, or moral norms, rather than epistemic norms per se. However, Carolina Flores and Elise Woodard (2023) argue for the view that based on epistemic norms and obligations regarding evidence-gathering, a rational agent is required to act and inquire within their biological and informational environment in a way that harvests evidence richness, rather than merely making correct inferences based on information already present in their mind.

The traditional view assumed that (1) belief is a passive state, whereas evidence-gathering is active; (2) epistemology only concerns norms governing mental states, not actions; and (3) the obligation to gather evidence is contingent upon the agent's practical goals. Flores and Woodard demonstrate that (*1) even processing evidence at hand requires active attention, option-structuring, and volitional inference; (*2) the traditional view cannot

consistently maintain that only norms governing mental states are epistemic, as it already accepts that abandoning a belief in light of overwhelming counter-evidence is an epistemic norm governing a mental act. Thus, epistemic rationality must have a broad scope encompassing norms governing both states and acts of inquiry; (*3) just as the norms of doxastic justification are independent of an individual's goals, the norms of evidence-gathering can also be truth-oriented norms, independent of personal preferences.

Flores and Woodard's core argument is that if there is an accepted procedure for the epistemic critique of an agent's act, then we have reason to believe that there are epistemic norms governing that act. Such an accepted procedure exists; therefore, we have reason to believe that epistemic norms governing evidence-gathering exist. [1]

- **Premise 1 (Justified Epistemic Blame):** Due to inter-subjective practices of epistemic critique, if an agent, in a significant matter, refuses to consult an accessible and efficient information source (such as AI search tools), relies solely on prior evidence, and subsequently falls into error, they may intuitively and appropriately be deemed deserving of epistemic blame and susceptible to charges of epistemic rashness or negligence.
- **Premise 2 (Blame Entails Normative Reasons):** Epistemic blame is justified only if the agent has a normative reason for x(a zetetic obligation to investigate, inquire, and evaluate evidence) and has refused to act accordingly without sufficient excuse.
- **Premise 3 (Application to AI Tools):** Advanced AI systems are currently among the prominent, efficient, and accessible resources for revealing errors, overlooked aspects of a problem, and alternative hypotheses.
- **Conclusion:** Therefore, the agent is rationally obligated (has most reason) to utilize available AI tools when forming beliefs on significant matters; deliberate refusal to do so constitutes a violation of norms of epistemic responsibility and warrants rational blame.

1. In everyday interactions, our established practices of epistemic criticism and blame reveal that we characteristically withhold epistemic trust from those who fail to inquire, gather superficial evidence, judge precipitously, lazily avoid further investigation, exhibit credulity, or display confirmation bias in their search. These practices imply that for the rational agent, merely "holding a belief or reaching a decision" is insufficient; rather, one is obligated to navigate one's epistemic environment, probe, compare, and enrich one's evidence base to avoid reliance on shallow or misleading data. Consequently, in significant decision-making contexts, indifference toward robust and accessible search tools—especially those capable of revealing higher-order evidence, serious objections, or cognitive blind spots—is simply incompatible with the prevailing norms of epistemic responsibility.

Argument 3: Diachronic Obligations and Evidence-Gathering

McWilliams (2014), writing from within the evidentialist camp, argues that an agent is obligated to reflect in certain cases (where complex evidence is involved). Because this reflection itself is a method for gathering new evidence (introspective evidence, intuitions, uncovering logical relations and hidden connections between beliefs), it constitutes a form of inquiry, and thus, the agent has an epistemic obligation to inquire. Inquiry is not merely "going to a laboratory" or "field research"; rather, mental review, inference, and the rearrangement of evidence (internal reflection) are themselves important modes of inquiry and evidence-gathering. Hall and Johnson (2018) demonstrate that if our epistemic end is to believe all and only true propositions, synchronic obligations are insufficient to achieve this goal. Epistemology must also include diachronic obligations – norms that guide our actions over time toward searching for and gathering fresh evidence.

- **Premise 1 (Insufficient Evidence and Evidence-Gathering):** For every proposition $p$regarding which the agent lacks decisive justification and the possibility of error is not ruled out, the agent has a pro tanto duty to search for and gather further evidence to reduce the risk of relying on incomplete data.
- **Premise 2 (Important Matters and All-Things-Considered Requirement):** In important matters, as the cost of error increases, the threshold for legitimate termination of inquiry rises. Under such conditions, the pro tanto duty to gather evidence can turn into an "all-things-considered" requirement, which cannot be overridden by minor cognitive or temporal costs. [1]
- **Premise 3 (Transmission of Obligation from End to Means):** If an agent is rationally obligated to achieve an optimal and justified judgment/decision regarding p, and performing action A is the most effective, least-costly method, or a necessary condition, for uncovering

1. In high-stakes situations, the rational threshold for relying on available evidence rises sharply, as the cost of potential error (asymmetric error cost) is significantly heavier than the marginal cognitive costs of time spent gathering further evidence. Therefore, in such circumstances, the weight that 'cost-of-gathering' considerations hold in the balance of practical rationality diminishes considerably. As a result, the duty to gather evidence—which was a *pro tanto* obligation in normal conditions—overcomes these marginal costs when faced with high risks, becoming an *all-things-considered* requirement for the agent (see Stanley, 2005; Fantl and McGrath, 2009). This inferential transition demonstrates that the use of AI tools to enhance evidence in critical decisions is not merely an optional choice, but an obligation stemming directly from zetetic norms linked to instrumental rationality.

unknown aspects and conflicting evidence regarding p, then the agent has a rational obligation to adopt means A.

- **Premise 4 (Relative Capacity and Neutrality of AI):** AI is a low-cost system that rapidly generates and formulates overlooked hypotheses, counter-reasons, and "higher-order evidence" concerning the agent's internal cognitive biases on a scale impossible for the individual mind.
- **Conclusion:** Therefore, in significant matters and conditions of uncertainty, a rational agent has a rational obligation to correctly utilize AI as a useful tool in the process of evidence-seeking and reasons-evaluation.

2) The Requirement of Practical Rationality

Rationality is not confined to the question of "what to believe"; it extends to the norms governing "how to think" and "how to gather evidence." Given the inherent limitations on human cognitive capacities and temporal resources, we are not required to deduce every logical consequence of our beliefs or to exhaustively gather all possible evidence for a given proposition. Consequently, the appropriate standard of rationality for human beings must reflect these cognitive bounds. Practical rationality, like theoretical rationality, is subject to these human limitations. Human agents naturally have multiple goals that cannot be realized simultaneously; thus, the rational selection among competing ends is an essential component of practical agency. Failure in this domain is distinct from epistemic failure in the realm of belief. Environmental conditions and the deliberative process play a decisive role in the rationality of decision-making. Many of our rational requirements (such as the need to seek further information or to resist temptation) exist precisely because of our cognitive and volitional limitations. Were we ideal rational agents, such "compensatory rational strategies" would be unnecessary. In short, rationality in the human sphere is a dynamic, multilayered system whose "requirements" fluctuate depending on the agent's epistemic situation, environmental opportunities, deliberative history, and even psychological traits. True rationality consists in the prudent management of these very limitations (see van Roojen, 2010).

- P**remise 1 (Practical Substantive Rationality):** Contrary to reductionist views (such as instrumental rationality), practical rationality is not limited to "finding the best means for given ends"; it encompasses the rational evaluation, harmonization, and prioritization of the ends themselves. An agent's fundamental goals constitute the structure of their life, their practical identity, and the allocation of their nonrenewable resources

(time, cognitive energy, and normative commitments). Therefore, a structural or macro-level error in the formulation of final ends may undermine the overall rationality of actions that are instrumentally effective relative to those ends.

- **Premise 2 (Deficiencies of Human Practical Rationality in Goal Selection):** Rational agents encounter persistent difficulties in selecting ends based on standard ethical theories:

    1. *From a Consequentialist Perspective:* Within the consequentialist tradition, rational goal selection depends on precise calculation of outcomes. Yet human capacities for such calculations are deeply bounded. Human actions and goals not only entail long-term, indirect consequences that are impossible for the agent to assess (Lenman, 2000), but in many cases, agents lack sufficient information regarding the weights of these consequences to calculate their expected value.
    2. *From a Deontological Perspective:* Assessing whether a "maxim" of action and its implicit end can be universalized requires a kind of coherence in intentions. Furthermore, its execution necessitates internal cognitive processes – such as precise maxim-formulation, unmixed motivation, and tracking volitional consistency – requiring the agent to meticulously avoid self-deception (Audi, 1988; Sticker, 2017).
    3. *From a Virtue-Ethical Perspective:* Under stress, time constraints, and motivational conflict, human agents often fail to act according to their virtues in particular circumstances, and the cultivation of virtue itself requires a protracted process of habituation (Miller, 2003).

- P**remise 3 (Compensatory Strategies):** Whenever a rational agent becomes aware of the limitations of their cognitive faculties, normative rationality requires the adoption of "compensatory strategies." An agent who is cognizant of the severe and predictable limitations of their cognitive, affective, and volitional capacities, yet persists in relying on errors resulting from them, is guilty of "epistemic and practical recklessness."

- **Premise 4 (Relative Efficacy of AI Systems):** Advanced AI systems possess processing and analytical mechanisms that can, to some extent, mitigate the aforementioned human limitations:

    1. *Addressing Incalculability in Consequentialism:* AI possesses a superior capacity to model complex systems, forecast multibranched scenarios, and calculate consequences beyond the reach of the human agent.

2. *Addressing Self-Deception in Deontology:* Algorithmic systems lack selfish motives, hidden interests, and psychological self-deception; they can, without being affected by these weaknesses, reveal formal contradictions, unprincipled exceptions, and tensions among conflicting commitments.
3. *Addressing Reduced Moral Sensitivity in Virtue Ethics:* Since human agents struggle to act virtuously under stress, time pressure, or motivational conflict – and given that habituation requires ongoing reinforcement – AI agents can prevent agents from falling into situational traps. This is achieved through predesigned selection models or timely warnings and reminders of moral commitments. Furthermore, AI can serve as a tool for habituation by recording, tracking, and analyzing the agent's behavioral patterns over time to provide personalized feedback for moral development.

- **Conclusion:** Therefore, assuming the availability of AI systems, a rational agent is rationally obligated to incorporate AI outputs into the process of forming, formulating, and reviewing their significant goals. Refusal to employ such tools can constitute rational negligence – an unjustified reliance on human faculties, the structural limitations of which are already established.

3) Challenges and Responses

It might be argued that if an agent seeks to utilize AI, (1) their autonomy will be undermined, potentially leading to a form of psychological/epistemic dependency that erodes the capacity for independent critical judgment over the long term; (2) their approach will drift away from a humanistic perspective toward a "mechanistic" one, falling prey to algorithmic bias; (3) the multiplicity of outputs from different AI systems may cause bewilderment and confusion; (4) one cannot trust the output due to AI's high error rate, lack of genuine understanding, or the opacity of its explanatory structure; (5) it creates a requirement for constant inquiry that leads to cognitive gridlock and practical paralysis, whereas practical rationality requires limits to inquiry; (6) it necessitates the abandonment of contextual/relational knowledge and lived experience; (7) it obscures the conditions for attributing responsibility when bad outcomes occur (a blameworthy agent could excuse her conduct by pleading, "I merely followed the rational requirement to consider the output of an approved system"); and (8) it creates a dilemma: (a) if the agent must independently evaluate and verify the AI output, they must possess expertise

in the related field (rendering the reliance on AI marginal), and (b) if the agent cannot verify it but is obligated to consider it, they have effectively engaged in blind acceptance rather than rational reflection.

In response, we contend that these challenges can be effectively defused once the nature of skillful human–AI engagement is properly conceptualized:

1. **On Autonomy:** The rational agent's reflective autonomy need not be undermined, provided that sovereign executive judgment and definitive normative endorsement in both theoretical and practical domains continue to reside strictly with the human agent.
2. **On Mechanistic Bias:** Because rational agents are aware of the architectural boundaries and potential biases of machine intelligence, they can treat the AI output as an algorithmic artifact, explicitly calibrating and distinguishing it from human and intersubjective considerations.
3. **On Confusion:** As the final decision-making authority remains with the agent, they can reflect upon multiple AI outputs and, as in the past, consult with qualified human experts.
4. **On Trust:** To evaluate the reliability of AI-generated output across diverse domains, an agent may adopt a tripartite strategy: first, utilizing multiple, independent AI systems to facilitate cross-system verification; second, subjecting these outputs to critical scrutiny and empirical testing prior to final adoption, thereby determining their context-specific reliability; and third, relying upon traditional, well-established epistemological procedures for knowledge acquisition as a corrective mechanism.
5. **On Inquiry Limits:** The rational obligation to consider AI output does not imply *ad infinitum* inquiry; it remains bound by the standard requirements of practical rationality.
6. **On Contextual Knowledge:** One can utilize this low-cost, high-speed technology without disparaging or discarding contextual/relational knowledge and lived experience.
7. **On Moral Responsibility:** Since the final selection and validation remain with the agent, responsibility for the decision resides with them; they cannot evade their moral burden.
8. **On the Expertise Dilemma:** To escape the dilemma, the agent should independently verify AI output where possible. In areas where they deem the AI output worthy of consideration but lack the expertise for full

verification, they should treat it as a "suspended" consideration (a pro tanto consideration) to be weighed until further investigation or consultation with competent human experts can resolve the matter.

4) Summary of the Normative Engagement

In light of the arguments presented, one may conclude that although contemporary AI systems and their prospective successors exhibit evident limitations and design constraints, these vulnerabilities do not undermine their advantages. They can be utilized – through training and the acquisition of requisite skills – to expand the rational agent's cognitive horizon. This allows the agent to leverage a vast array of information, knowledge, and reports of human experience, thereby minimizing human cognitive biases, deficits, and errors. If an AI system merits an agent's trust within a given domain, the agent may be rationally required to appropriately deploy their skillfully generated outputs as a pro tanto or prima facie consideration in their significant theoretical and practical decisions (whether for belief formation or goal selection).

Descriptive Engagement with AI

The normative analysis established that, under appropriate epistemic and practical conditions, rational agents may be required to engage critically with AI-generated outputs rather than exclude them from significant deliberation. The descriptive question that follows is what happens when such engagement becomes sustained and integrated into human practices. The increasing incorporation of artificial intelligence into everyday activities is transforming the conditions under which individuals perceive, deliberate, act, and understand themselves. Earlier approaches to technology largely treated technological systems as instruments for realizing purposes independently formed by human agents. Contemporary AI complicates this instrumental model by generating information, interpretations, alternatives, and recommendations that enter directly into ongoing processes of human cognition and action. Phenomenological approaches have long shown that technologies can shape how human beings encounter and interpret the world rather than merely serving as neutral means (Ihde, 1990). AI intensifies this mediating function by generating new symbolic and informational possibilities within interaction itself.

Recent accounts of artificial agency and human–AI interaction therefore suggest that the relation between humans and AI cannot be adequately understood through a simple subject–object distinction. Noller (2024)

conceptualizes AI as extending human agency through an ongoing relation between human and artificial performances of intelligence, while research on human–AI collaboration indicates that agency can become distributed across processes of interaction, contribution, and attribution (Holter et al., 2026). From a broader perspective, technological mediation shapes the relations among subjects, objects, and practices rather than merely intervening between them (Verbeek, 2005). These developments become especially significant when AI moves beyond task assistance and enters the processes through which individuals formulate intentions, evaluate alternatives, construct narratives, and reflect upon themselves.

The descriptive question is therefore not whether AI replaces human agency or produces an artificial form of human selfhood. It is how sustained AI mediation reorganizes the conditions through which human beings act and experience themselves as acting and self-interpreting subjects. This section examines this transformation through four interconnected stages: it first considers the transformation of the field of human agency, then the reconfiguration of human selfhood, before moving from technological mediation toward relational constitution and finally specifying these changes as an ontological reconstitution of the human subject.

1) AI and the Transformation of the Field of Human Agency

Artificial intelligence transforms the field of human agency not merely by increasing the efficiency of existing tasks, but by entering the processes through which possibilities for action are identified, organized, evaluated, and rendered intelligible. In a conventional instrumental conception of technology, the subject forms an intention, selects a means, and employs it to realize an already determined purpose. Contemporary AI complicates this structure by generating alternatives, organizing information, identifying patterns, formulating interpretations, and anticipating possible consequences. It can therefore participate in shaping the practical possibilities from which human intentions and actions emerge. The transformation concerns not only how efficiently humans act, but also the conditions under which intentions become actionable.

This development marks a shift from technological assistance in execution toward participation in possibility formation. Individuals may approach AI with open-ended questions, incompletely formulated problems, or indeterminate objectives, after which AI can generate alternative formulations, relevant information, possible courses of action, and interpretive frameworks. Noller's (2024) account of artificial agency is relevant because it moves beyond a strict opposition between human and artificial agency and emphasizes how

artificial systems can become involved in extending and realizing human activity. Similarly, technological mediation can shape the relations through which the world becomes available for human action and interpretation rather than merely functioning as an intermediary between subject and world (Verbeek, 2005). AI intensifies this mediating function by producing new possibilities within the activity itself.

This transformation requires distinguishing decision ownership from possibility formation. A human agent may retain final decision-making authority while no longer independently determining the informational and practical field from which alternatives are drawn. AI can identify options, prioritize information, reveal patterns, and formulate possible courses of action that subsequently become objects of human deliberation. The decision remains attributable to the human agent, while the field of deliberation may be technologically structured. Holter et al. (2026) similarly emphasize that human–AI interaction can distribute agency across processes of interaction, contribution, and attribution. Agency should therefore be examined not only at the moment of choice but also through the preceding processes that determine what becomes available for perception, evaluation, and action.

The result is a technologically structured possibility space. Human agency continues to involve perception, problem formulation, deliberation, intention, action, and evaluation, but these processes increasingly unfold within environments partly organized by computational systems. AI can influence what becomes salient, which alternatives appear comparable, how problems are framed, and which consequences become relevant. This does not require attributing human-like intentionality or consciousness to AI. As Dung (2025) argues, the philosophical challenge concerns the relationship between artificial systems and the human capacities through which action and agency are understood. The central issue is therefore not whether AI has acquired human agency, but how it transforms the field within which human agency is exercised.

Within this altered field, intention itself can become relationally formed. Human purposes may be specified, revised, and rendered more determinate through iterative engagement with computationally generated possibilities. A researcher may begin with a broad question, receive alternative conceptual framings, reject some, develop others, and return to AI for further refinement. The resulting intention remains human in its experiential and practical locus, while its articulation may emerge through an interaction in which computational contributions shape the available conceptual possibilities. This can be described as intentional co-formation without implying that AI possesses intentions equivalent to those of human subjects. Floridi (2025) helps explain this process by emphasizing AI's capacity to transform informational

and semantic environments, thereby affecting not only what individuals do but also the conditions under which purposes become intelligible and actionable.

Agency consequently appears less as an isolated capacity exercised at the final moment of choice and more as a temporally distributed process. What a person does depends partly on how a problem was framed, what information became salient, which alternatives became visible, how they were interpreted, and how intentions were subsequently revised and translated into action. AI can enter several of these stages, creating an ongoing interaction rather than a single technological intervention. This does not imply a transfer of agency from humans to machines. Human and computational contributions remain asymmetrical. The human subject experiences, evaluates, commits, and acts, whereas AI generates and transforms computational possibilities. Yet these contributions can become structurally interconnected within sustained activity, producing a relationally organized form of human agency.

The transformation is therefore best understood neither as the replacement of human agency nor as the simple enhancement of pre-existing capacities. It is a reconfiguration of the temporal, informational, and relational conditions under which human agency is exercised. AI becomes significant because it can participate in forming the possibilities through which human intentions acquire determinate form. Human agency thus remains human, while the field in which it is articulated becomes increasingly shaped through sustained interaction with computational systems.

2) AI and the Reconfiguration of Human Selfhood

If the transformation of human agency concerns changing conditions of action, the reconfiguration of selfhood concerns changing conditions of self-interpretation. Selfhood is not reducible to biological identity, psychological continuity, or stable personality traits. It also involves the reflexive capacity through which individuals interpret experiences, attribute actions and commitments to themselves, and establish continuity between who they have been, who they are, and who they may become. Ricoeur's ([1990] 1992) distinction between sameness and selfhood is crucial here. Sameness concerns continuity of characteristics, whereas selfhood involves an ongoing interpretive relation through which a person recognizes themselves across temporal and experiential difference. The increasing incorporation of AI into practices of remembering, describing, interpreting, and anticipating therefore affects not only what individuals know about themselves but also the conditions through which they become intelligible to themselves.

AI's significance for selfhood can be understood through algorithmic

reflexivity. Unlike a conventional mirror, AI systems generate computationally organized representations through classification, comparison, prediction, summarization, and interpretation. They can identify recurring patterns in personal writings, summarize previous interactions, infer preferences, and formulate descriptions of behavioral tendencies. What is returned to the individual is therefore not a neutral reflection but a technologically mediated representation of the self. AI does not simply reveal an already constituted self. It can produce representations that subsequently become materials for self-interpretation, extending the broader insight that technologies can shape how subjects encounter and understand themselves (Verbeek, 2005).

This becomes particularly significant when AI-generated representations enter the narrative organization of personal experience. Human self-understanding is partly narrative because experiences acquire coherence through their relation to memories, commitments, actions, and anticipated possibilities. Ricoeur's ([1990] 1992) account of narrative identity emphasizes that selfhood develops through the interpretation of heterogeneous experiences within an intelligible temporal configuration. AI can participate in this process by organizing autobiographical materials, identifying recurring themes, summarizing personal records, and generating interpretations of patterns across previous interactions. Such representations can then be accepted, rejected, revised, or incorporated into an individual's developing self-description. AI therefore becomes an additional element within the interpretive activity through which individuals understand themselves.

This process also transforms the relation between self-recognition and self-description. An individual may experience a recurring tendency without having conceptualized it as a significant feature of the self. When AI identifies and articulates such a pattern, something previously implicit may become available for explicit reflection. The individual can then recognize the description as accurate, incomplete, or misleading and revise it accordingly. AI thus becomes part of a reflexive process through which the subject observes, evaluates, and redescribes itself. The significance lies not in attributing self-consciousness to AI, but in recognizing its increasing involvement in the conditions through which human self-consciousness is exercised. As such representations become personalized and temporally continuous, they may become embedded in ordinary practices of self-monitoring and self-interpretation.

The same dynamic extends from the interpretation of the past to the imagination of possible futures. Markus and Nurius's (1986) concept of possible selves emphasizes that individuals understand themselves partly through imagined states they might become, hope to become, or fear becoming. AI can participate in this anticipatory dimension by generating

scenarios concerning possible decisions, careers, behaviors, or forms of personal development. By reorganizing information about the past and present into representations of possible futures, AI can contribute to the imaginative space within which individuals interpret who they might become. Selfhood can therefore be understood as temporally distributed across remembered experience, present interpretation, and anticipated possibility.

Yet AI's involvement does not establish an equivalence between human and artificial subjectivity. Human beings experience, interpret, appropriate, and live the significance of these representations, whereas AI produces computational representations without becoming a first-person subject of the life represented. This distinction remains compatible with accounts of extended cognition that recognize the functional integration of external systems without collapsing the difference between human experience and technological operation (Clark & Chalmers, 1998).

The descriptive significance of AI for selfhood therefore lies not in the replacement of human subjectivity but in the reorganization of its reflexive conditions. When computational representations become recurrently integrated into remembering, self-description, interpretation, and anticipation, AI becomes structurally involved in how individuals recognize and narrate themselves across time. Human selfhood remains grounded in lived, embodied, and first-person experience, while the processes through which that selfhood is interpreted become increasingly relational and technologically mediated. AI thus contributes to a reconfiguration of selfhood by becoming part of the reflexive processes through which human beings understand not only what they have done, but who they are and who they may become.

3) From Technological Mediation to Relational Constitution

The transformation of human agency and selfhood points toward a deeper change in the human–AI relation, namely a movement from technological mediation toward relational constitution. If AI participates not only in executing human purposes but also in forming possibilities, interpretations, and self-understandings, it can no longer be adequately described as an external intermediary. Technologies have long shaped how human beings perceive and engage with the world rather than functioning as neutral instruments (Ihde, 1990). Contemporary AI intensifies this condition because generative systems can produce new informational, symbolic, and practical possibilities within the activities they mediate. The central question is therefore when technological mediation becomes sufficiently integrated into a practice that the technological relation forms part of its constitutive organization.

This distinction can be clarified by differentiating instrumental, interactive, and generative technologies. Instrumental technologies primarily serve purposes established independently of them, whereas interactive technologies respond to and modify ongoing activities. Generative AI introduces a further transformation by producing formulations, alternatives, classifications, interpretations, and projections that can influence how an activity itself is organized. Its outputs may therefore affect not only how an intention is executed but also how that intention is initially articulated. Verbeek's (2005) account of technological mediation is relevant because technologies shape the relations through which humans encounter and interpret the world. With generative AI, this mediating function becomes dynamic and recursive, as the system responds to human activity while producing new elements that alter its subsequent course.

The resulting relation is better understood as participation rather than mere mediation. Participation does not imply that AI becomes a human subject or acquires human intentionality. Rather, computational operations become dynamically integrated into an unfolding human practice. A person may formulate an objective, receive computationally generated possibilities, accept or reject them, revise the original objective, and incorporate the resulting output into subsequent action. Human and computational operations thus remain ontologically distinct while becoming functionally interconnected. This structure can be described as coupled agency, in which human purposes and computational operations become dynamically coordinated within sustained activity without becoming equivalent forms of agency. Such an account is consistent with approaches that understand human–AI activity through interaction, contribution, and distributed attribution rather than the isolated action of a single agent (Holter et al., 2026).

The significance of this coupling becomes clearer when agency and selfhood are considered together. Human actions are remembered, interpreted, attributed, and incorporated into subsequent understandings of the self. When AI participates in forming possibilities preceding action, its contributions can therefore enter later processes of self-interpretation. Computationally generated possibilities can shape action, action can become an object of self-interpretation, and revised self-understanding can influence subsequent intentions and further interactions with AI. The significance of this recursive structure does not lie in attributing subjectivity to AI, but in recognizing that computational contributions can become embedded within the temporal continuity through which human agency and selfhood develop. Ricoeur's ([1990] 1992) account of narrative identity is relevant because selfhood is constituted through the ongoing interpretation and integration of actions and experiences across time.

This recursive integration requires distinguishing influence from

constitution. Not every technological influence constitutes an ontological transformation. Human beings have always depended upon technologies, and the modification of behavior by a technological artifact does not by itself establish that the technology has become constitutive of a practice. Constitution requires stronger integration. The technology must repeatedly enter the organization of an activity, modify relations among its stages, and become part of the conditions through which the activity is performed and subsequently interpreted. The relevant distinction is therefore between episodic mediation and sustained relational integration.

This qualification is essential because a single AI recommendation or occasional interaction does not by itself reconstitute human agency or selfhood. The constitutive claim becomes descriptively plausible when AI-mediated operations acquire continuity within a practice and later stages depend upon, respond to, or presuppose earlier computational contributions. Under such conditions, AI participates in reorganizing the practical and interpretive relations through which action acquires its form and significance. This understanding is also compatible with the extended-mind tradition, which shows how external technological resources can become integrated into cognitive processes under appropriate conditions of functional organization (Clark & Chalmers, 1998).

Relational constitution therefore describes a mode of human–AI involvement in which human capacities remain distinct while their exercise becomes increasingly organized through sustained technological interaction. It does not imply a fusion of human and artificial subjects or an ontological equivalence between them. Rather, it identifies a structured interdependence between embodied human experience and computational generation. What changes is the relational architecture within which human beings act, interpret their actions, and develop self-understanding. Under conditions of sustained integration, technological mediation becomes part of the constitutive conditions through which human agency and selfhood are dynamically organized. AI therefore matters not because it replaces the human subject, but because it becomes increasingly embedded in the relations through which that subject acts and understands itself.

4) The Ontological Reconstitution of the Human Subject

The preceding analysis clarifies what is meant by the ontological reconstitution of the human subject. The claim is neither that artificial intelligence replaces the human subject nor that AI becomes a human-like subject. Rather, the transformation concerns the relational conditions under

which human beings continue to exist as acting, knowing, and self-interpreting subjects. A classical instrumental conception assumes a relatively clear distinction between the human agent and the technological object, with the subject forming an intention and technology serving as an external means for its realization. AI-mediated activity complicates this separation because computational systems increasingly participate in the formation of possibilities, organization of action, interpretation of information, and incorporation of experience into self-understanding. The human subject therefore remains human, while the conditions through which human subjectivity is enacted become increasingly relational.

This transformation should not be understood as a movement from autonomous subjectivity to technological determination. More precisely, it involves a shift from a subject conceived as primarily independently constituted toward a subject whose modes of acting and self-interpreting are increasingly relationally constituted. Relational constitution does not eliminate the distinctive mode of human existence or establish ontological symmetry between humans and AI. Human beings remain embodied, temporally situated, experientially conscious subjects whose activities unfold within a lived world. AI systems, by contrast, remain computational entities dependent upon algorithms, data, technical infrastructures, and material systems. The relational character of human–AI existence therefore preserves this ontological difference while identifying a transformation in the conditions under which distinctively human capacities are exercised.

Ontological reconstitution can consequently be understood as a transformation in the conditions through which human beings become and remain agents and self-interpreting subjects. The significance of reconstitution lies in its systemic rather than isolated character. AI increasingly enters several interconnected dimensions of human existence: it participates in the practical organization of action by generating and structuring possibilities, in the epistemic organization of experience by filtering and reorganizing information, and in the reflexive organization of selfhood by producing representations through which individuals interpret their actions, memories, characteristics, and possible futures. These dimensions are mutually connected. What a person can do affects what the person knows, what the person knows influences the interpretation of previous actions, and such interpretations contribute to subsequent intentions.

The first dimension is therefore practical reconstitution. AI increasingly participates in the formation and organization of possibilities that precede and inform action. Human action may consequently unfold within technologically structured fields in which computational systems contribute to the generation

and transformation of alternatives. The second dimension is epistemic reconstitution. AI does not merely provide information after a human has independently established a way of understanding a problem. It can also influence which information becomes salient, how it is organized, and which interpretations become available for consideration. Human knowing is thus increasingly situated within informational environments partly structured through computational operations. The third dimension is reflexive reconstitution. By generating representations of patterns, preferences, experiences, and possible futures, AI can become involved in the processes through which individuals interpret themselves and establish continuity between past experience, present identity, and anticipated possibilities.

These three dimensions of agency, knowledge, and selfhood should therefore be understood as interconnected rather than isolated. Human beings act within fields of possibility, interpret the conditions and consequences of their actions, and incorporate those interpretations into an ongoing understanding of themselves. AI can become integrated across these interconnected processes. Its significance does not derive from possessing the same capacities as the human subject, but from becoming part of the relational environment through which those capacities are exercised and connected.

The ontological reconstitution of the human subject thus describes a transformation in the architecture of human existence rather than the emergence of an artificial human subject. The human remains the embodied and self-interpreting locus of experience, while the conditions of agency, knowledge, and self-relation increasingly include technologically generated operations. What is reconstituted is therefore not human identity itself, but the relational configuration through which human beings act, understand, and interpret themselves. AI does not replace the human subject or become its ontological equivalent. It transforms the relational conditions through which the human subject exists as an acting, knowing, and self-interpreting being.

5) Summary of the Descriptive Engagement

Human intentions and decisions are formed, revised, and determined through an ongoing process of reciprocal interaction with artificial intelligence. Given AI's algorithmic character, this interaction transforms the ways in which agents narrate themselves, under the influence of structured, computational representations. Through this transformation, the subject – as a knowing, acting, and self-interpreting being – continues to exist within a fundamental relation. In other words, by technologically configuring new possibilities that precede action, structuring the informational environment, evidence, and the

framing of problems, and shaping human recognition, self-monitoring, and the formation of narrative identity over time, artificial intelligence turns the human being into a subject whose modes of action and self-interpretation are organized as "relationally constituted."

Conclusion

The rational obligation to utilize AI (the first thesis), combined with the transformation of the agent's practical identity (the second thesis), can predictably pave the way for more numerous, burdensome, or complex rational or moral duties. This is because rational or moral obligations are dependent on the agent's practical identity, as well as their epistemic and intentional status. When these states are elevated or enhanced, the agent's duties naturally become more extensive or demanding. To examine this consequence more precisely, consider the following scenario:

- **Agent (*A*):** Sarah, a major philanthropist and director of a family charitable foundation.
- **Sarah's Goal at t_1:** "Allocating a large financial endowment (tens of millions of dollars) to reduce poverty, malnutrition, and school dropout rates in a marginalized, multinational geographical region simultaneously suffering from ecological crises (drought) and social instability." (This goal possesses the properties of "high-stakes outcomes," "high probability of resource waste and intervention failure," and "incalculable complexity for the individual human mind.")

Phase One: t_1 – The Rational Obligation to Critically Consult the System

(*Realization of Thesis 1: Rational obligation arising from higher-order evidence, the mitigation of cognitive biases, and the epistemic duty of inquiry*)

- **Status at t_1:** Influenced by direct moral intuitions, emotionally charged narratives, and empathy-driven biases, Sarah is inclined to spend the entire budget on building schools and sending immediate relief packages. She perceives herself as a "traditional philanthropist with a clear moral conscience and good intentions."
- **Introduction of AI System (*S*):** System *S*(an AI capable of development modeling and massive socio-ecological data analysis, with the ability to predict resilience, local market flows, and unintended side effects over a ten-year horizon) is available.

- **Rational Obligation:**

1. *Higher-Order Evidence:* Assume that studies in behavioral economics and applied moral philosophy show that philanthropists suffer from severe cognitive biases, such as "scope neglect" and the "identifiable victim effect." Sarah possesses evidence indicating that her individual deliberative process is highly fallible in predicting systemic outcomes.
2. *Zetetic Obligation:* Given that wasting these resources means losing the opportunity to save the lives of hundreds of children, possessing "good intentions" alone does not satisfy practical rationality. Rationality requires Sarah to make sufficient use of the available data before making a significant decision.

- **Conclusion of Phase One:** Therefore, consulting S is not merely a preference or a permissible act; it is rationally mandatory. The output of Sprovides a pro tanto reason: the model predicts that building schools without securing water safety and reforming agricultural livelihood cycles will lead to rural depopulation and leave those schools largely unused within three years. Sarah critically evaluates the output and adjusts the budget according to the model.

Phase Two: The Transition from $t_1$to t_2 – Continuous Interaction and the Reconstruction of Practical Identity

(Realization of Thesis 2: Relational constitution, algorithmic reflexivity, and the transformation of self-narrative)

- **Continuous and Recursive Interaction:** Over several years, Sarah works consistently with AI analytical platforms for all development projects, humanitarian crisis forecasting, and intervention strategies:

1. *Possibility Formation:* Novel methods of philanthropy (such as conditional cash transfers or algorithmic redesign of food supply chains) are discovered and formulated in interaction and co-creation with AI's computational possibilities.
2. *Algorithmic Reflexivity:* The system shows Sarah how her past charitable preferences were influenced by a paternalistic gaze. The AI presents her with a reflective graph of her "unconscious valuation patterns."

- **Transformation of Practical Identity at t_2 (Reconfigured Practical Identity):** The description under which Sarah deems her actions meaningful undergoes a fundamental transformation:

- At t_1, her practical identity was: "A compassionate and generous benefactor who bestows her personal wealth upon the suffering based on moral sentiments and individual autonomy."
- At t_2, her practical identity has become relationally constituted: She now understands herself as: "A strategic and moral figure within a hybrid (human–machine) decision-making network, whose mission is the radical engineering of justice structures, the resilience of vulnerable communities, and the systemic optimization of scarce life-sustaining resources." Her narrative selfhood is no longer that of a "solitary savior," but rather an agency coupled with data analysis systems.

Phase Three: t_3 – The Emergence of New, More Complex, and Burdensome Rational and Moral Duties

(New rational or moral duty stemming from the fulfillment of rational obligations and a new practical identity)

- **Emerging Crisis at t_3:** Sarah's foundation intends to implement a large-scale nutrition and treatment project across diverse climates. The AI optimization models suggest an option that maximizes overall health indicators at a macro level. However, a native minority ethnic group on the region's periphery, due to physiological and structural differences, benefits the least from this proposal and is effectively marginalized.
- **The New Moral and Rational Duty at t_3:** Sarah now faces a highly complex and binding duty:

> "The duty to re-evaluate and reform AI models in favor of intergenerational justice and minority rights (algorithmic justice auditing), to resist blind computational utilitarianism, and to systematically redesign the model to guarantee the human dignity of 'invisible' groups, even if it slightly reduces the overall statistical efficiency index."

- **Why did this duty not exist before t_1 and would not have emerged without this path?**

1. *Regarding Epistemic Status:* A traditional philanthropist at t_1 remains in a state of blameless ignorance regarding these types of systemic biases and hidden distributive consequences; they lacked the capacity to process the cross-impacts of ethnic and data variables, and simply sending food fulfilled their duty. By t_3, however, due to long-term interaction with the

systems, Sarah possesses the epistemic mastery and tools to uncover "algorithmic blind spots." When epistemic capacity increases, the scope of moral responsibility broadens: she can no longer claim ignorance of the model's hidden distributive consequences.

2. *Regarding Practical Identity and Role Obligations:* Moral duties are partially rooted in the agent's practical identity. At t_1, if Sarah had failed to notice a complex side effect on a native minority, her moral identity would not have been compromised; she was merely a "single philanthropist providing direct alms." But for Sarah at $t_2$and t_3, violating this fair audit means a direct violation of her new practical identity. She has constructed her identity based on being a "regulator of structural justice and an operator of complex systems"; abandoning this duty at t_3 would cause her personality coherence and role commitments to collapse from within.
3. *Counterfactual Check:* If Sarah had not submitted to the rational obligation of consulting the AI at $t_1$and had remained within traditional charitable methods:

- Her epistemic status would not have reached the horizon of structural analysis and second- or third-order consequences;
- Her practical identity would not have been rearranged as a subject with relational constitution;
- Consequently, at t_3, no one (not even her own conscience) would have held her responsible for failing to perform an algorithmic audit of distributive justice.

This example clearly demonstrates how the rational obligation to utilize AI, and the transformation of the agent's practical identity through long-term interaction, can predictably pave the way for more numerous, burdensome, or complex rational or moral duties for the rational agent.

▣ Conflict of Interest

The authors declare no competing interests.

References

Audi, R. (1988). Self-deception, action, and will. *Erkenntnis*, *28*(2), 247–265. https://doi.org/10.1007/BF00166408

Clark, A., & Chalmers, D. J. (1998). The extended mind. *Analysis*, *58*(1), 7–19. https://doi.org/10.1111/1467-8284.00096

Dung, L. (2025). Understanding artificial agency. *The Philosophical Quarterly*, *75*(2), 450–472. https://doi.org/10.1093/pq/pqae010

Fantl, J., & McGrath, M. (2009). *Knowledge in an uncertain world*. Oxford University Press.

Flores, C., & Woodard, E. (2023). Epistemic norms on evidence-gathering. *Philosophical Studies*, *180*(9), 2547–2571. https://doi.org/10.1007/s11098-023-01987-9

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. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.5135645

Hall, J., & Johnson, A. (2018). The epistemic duty to seek more evidence. *American Philosophical Quarterly*, *55*(2), 173–186.

Holter, S., Moruzzi, C., & El-Assady, M. (2026). Towards agency in human-AI collaboration. *IEEE Computer Graphics and Applications*, *46*(1), 13–25. https://doi.org/10.1109/MCG.2025.3623892

Ihde, D. (1990). *Technology and the lifeworld: From garden to earth*. Indiana University Press.

Lenman, J. (2000). Consequentialism and cluelessness. *Philosophy & Public Affairs*, *29*(4), 342–370. https://doi.org/10.1111/j.1088-4963.2000.00342.x

Markus, H., & Nurius, P. (1986). Possible selves. *American Psychologist*, *41*(9), 954–969. https://doi.org/10.1037/0003-066X.41.9.954

McWilliams, P. (2014). Evidentialism and epistemic duties to inquire. *Pacific Philosophical Quarterly*, *95*(2), 263–282. https://doi.org/10.1111/papq.12028

Miller, C. (2003). Social psychology and virtue ethics. *Journal of Ethics*, *7*(4), 365–392. https://doi.org/10.1023/A:1026040810355

Noller, J. (2024). Extended human agency: Towards a teleological account of AI. *Humanities and Social Sciences Communications*, *11*, Article 1338. https://doi.org/10.1057/s41599-024-03849-x

Palmira, M. (2024). Higher-order evidence and the duty to double-check. *Noûs*, *58*(3), 799–824. https://doi.org/10.1111/nous.12480

Ricoeur, P. (1992). *Oneself as another* (K. Blamey, Trans.). University of Chicago Press. (Original work published 1990)

Stanley, J. (2005). *Knowledge and practical interests*. Oxford University Press.

Sticker, M. (2017). When the reflective watch-dog barks – Conscience and self-deception in Kant. *Journal of Value Inquiry*, *51*(1), 85–104. https://doi.org/10.1007/s10790-016-9554-y

van Roojen, M. (2010). Moral rationalism and rational amoralism. *Ethics*, *120*(3), 495–525. https://doi.org/10.1086/652298

Verbeek, P.-P. (2005). *What things do: Philosophical reflections on technology, agency, and design*. Pennsylvania State University Press.