Philosophy of AI
Vol. 1, 2025, 107-110
10.18716/ojs/phai/2025.11993

EDITORIAL OF THE INAUGURAL VOLUME OF "PHILOSOPHY OF AI"

Guido Löhr

Vrije Universiteit Amsterdam, Philosophy, Amsterdam, NL

It is with great pleasure that we welcome readers to the inaugural issue of *Philosophy of AI*, the first diamond open-access journal dedicated to the philosophical study of artificial intelligence.

Philosophy of AI is the official double-blind journal of the Society for the Philosophy of AI, both of which were founded at the biannual PhAI Conference in Erlangen in December 2023.

Since this journal is also supposed to be a platform for our society, we first review the activities of the society this year. It has been a very productive year for us. Not only did we publish the first volume of our journal this year, but we also installed a new website (https://phai.ac/) and a mailing list (run by Ian Robertson). Finally, we organized the 6th biannual PhAI conference in Amsterdam this Fall. Our keynote speakers were Mona Simion, Emily Sullivan, and Markus Kneer.

If you would like to become a member or be notified about upcoming events, please sign up for our mailing list. At present, the society is deliberately run non-bureaucratically, which also means that there are no membership fees. Further information can be found at https://phai.ac/membership/. In the future, we plan to develop a platform that will further strengthen this network and make its many connections visible and accessible.

Back to the journal.

As mentioned, the journal is run in the spirit of the Open Access movement. This means that publishing with us is entirely free of charge. We welcome submissions on philosophical questions related to artificial intelligence, ranging from ethics and political philosophy to philosophy of mind, epistemology, metaphysics, and philosophy of language, with an analytic orientation broadly construed.

The reason Vincent Müller and I founded this journal is that artificial intelligence has become one of the most intensively discussed topics across academia, industry, and public discourse. Alongside technical developments, there is a growing demand for philosophical reflection on both the normative implications of AI.

Much of this discussion is taking place within philosophy. Over the past few years, the institutional impact of AI on the discipline has been striking: new professorships, research groups, centers, and degree programs explicitly devoted to the philosophy of AI have emerged at leading research institutions worldwide. Philosophy of AI is no longer a niche topic; it is rapidly becoming a central area of philosophical inquiry.

Despite this growth, there has until recently been no journal devoted exclusively to philosophy of AI. Existing venues—such as *Minds and Machines*, *Philosophy & Technology*, *AI & Society*, or *AI and Ethics*—play an important role in the landscape, but their scope either extends well beyond AI or beyond philosophy. None of them focuses specifically on philosophical work that fits the short and precise title *Philosophy of AI*.

Our aim is to provide a clear disciplinary home for this emerging field.

Equally important is our commitment to open access. At present, no journal in analytic philosophy of AI operates fully open access. *Philosophy of AI* responds to this gap by offering Gold Open Access publication under a CC-BY license, with no article processing

charges for authors and no fees for guest editors of special issues.

Starting a journal has been exciting and fun, but also difficult, and running an independent, open-access journal certainly has its challenges. However, many of these problems are now fixed, and the journal is looking toward a promising future.

Our opening volume brings together contributions that exemplify the range and ambition of the journal's mission.

The issue opens with Cappelen, Goldstein, and Hawthorne's (2025) "AI Survival Stories: a Taxonomic Analysis of AI Existential Risk," which brings much-needed structure to discussions of AI-induced extinction. Instead of defending or rejecting doomsday claims directly, the authors decompose standard arguments into two premises: that AI systems will become extremely powerful, and that such systems would then destroy humanity. By mapping ways in which either premise might fail, they develop a taxonomy of "survival stories," each associated with distinct empirical challenges and policy responses. The result is a framework that makes explicit where disagreements about existential risk really lie, and how different assumptions motivate different practical interventions.

This target article was discussed in responses by Rory Svarc (2025) *"Defending Alignment: A Commentary On 'AI Survival Stories',"* Aksel Sterri and Peder Skjelbred (2025) *"Automation and its Discontents: Leveraging Automation to Safeguard Humanity"*, Leonard Dung (2025) *"Estimating the probability of AI existential catastrophe: Converging on the answer from opposite ends"*, and Kate Vredenburgh (2025) *"AI survival stories, types of risk, and the precautionary principle."* Cappelen, Goldstein, and Hawthorne (2025b) responded to their critics in their *"AI Survival Stories - Responses to Critics."*

In their analysis, Cappelen, Goldstein, and Hawthorne presuppose a relatively stable conception of "human survival," whereas Andrea Sauchelli's (2025) "Artificial Intelligence, Ontology, and Existential Risks" questions that presupposition itself. Sauchelli asks what ontological assumptions about ourselves underwrite contemporary existential-risk discourse, and whether those assumptions can be taken for granted. If future AI systems were to count as part of "us," or as legitimate successors to humanity, then familiar formulations of existential risk would require substantial revision. By linking AI risk to theories of personal and collective ontology, the paper shows that debates about extinction are inseparable from deeper questions about identity and continuity.

A complementary shift in perspective is offered by Dina Babushkina (2025) in "AI, decisions, and the reasons to believe: an ethics-through-epistemology approach." Treating AI as a cognition technology, Babushkina reframes responsibility not in terms of AI agency or metaphysical blameworthiness, but in terms of the epistemic conditions under which human decisions are made. The paper shifts focus from actions to decisions, from imputability to harm mitigation, and from ontological to epistemic criteria. According to Babushkina, responsible AI use requires critical evaluation of the reasons for belief delivered by AI systems, especially as they acquire increasing epistemic authority. Responsibility, here, is fundamentally a matter of managing belief and justification in AI-mediated cognition.

The final cluster of papers is a part of an effort to promote philosophy of language on AI via an ongoing "topical collection," edited by Mitch Green, Jan Michel, and me.

In "Babbling stochastic parrots? A Kripkean argument for reference in large language models," Steffen Koch (2025) challenges the increasingly popular claim that LLMs lack semantic competence. Drawing on causal–historical theories of reference, Koch argues that LLMs can successfully refer to objects and kinds by inheriting reference from their training data via a reference-sustaining mechanism. This account aims to show that meaningful language use does not require internal understanding or intentions.

Mitchell Green's (2025) "Large Language Models and the Varieties of Meaning" responds by rejecting the causal-theoretic strategy as both controversial and limited in scope. Green proposes instead an Austin-inspired account on which LLMs count as using language insofar as they perform rudimentary phatic acts. Meaning, on this view, is grounded in minimal communicative practice rather than in reference-fixing mechanisms. The exchange is continued in Koch's (2025b) "What does it take to establish reference in LLMs?

Kripke vs. Austin *(Response to Green)*" which defends the Kripkean approach and raises objections to the speech-act alternative.

Taken together, these papers exemplify the kind of philosophical work this journal aims to promote: work that does not treat AI as a mere occasion for applied ethics or speculative metaphysics, but as a site of genuine conceptual disruption.

These contributions set a high standard, and we are grateful to the authors, reviewers, and members of our editorial board whose care and intellectual generosity made this issue possible.

Looking forward, we invite submissions that challenge disciplinary boundaries, question entrenched assumptions, and offer new conceptual tools for understanding the role of AI in contemporary life. We particularly encourage work that bridges philosophical theory with technical, social, and political contexts. As AI continues to transform not just what we can do, but what we can *be*, philosophical reflection becomes indispensable.

The success of a diamond OA journal depends on a community committed to shared scholarly values: intellectual rigor, openness, inclusivity, and mutual support. We hope *Philosophy of AI* can provide a forum for such a community.

We are grateful to everyone who helped and supported us, in particular our support team in Cologne: Joao Martins, Eric Eggert. We thank the "Fachinformationsdienst Philosophie", funded by the German Research Council, for their support (https://philportal.de).

We also thank our editorial assistant, Eleonora Catena, and our many excellent editors: Björn Lundgren (Friedrich Alexander Universität Erlangen-Nürnberg, FAU), Katsunori Miyahara (Hokkaido University), Sven Nyholm (Ludwig Maximilian University, Munich, LMU), Jakob Ohlhorst (RWTH Aachen), Lucy Osler (Cardiff University), and Adrian Yee (Hong Kong Baptist University).

We also thank our advisory board for putting trust in our project: Colin Allen (UC Santa Barbara), Cameron Bruckner (University of Houston), Herman Cappelen (HKU, Editor-in-Chief of Inquiry), Catarina Dutilh Novaes (VU Amsterdam), Mitchell Green (University of Connecticut), Jan Michel (Heinrich Heine University Düsseldorf), Rachel Sterken (HKU), Andrea Sauchelli (Lingnan University)

We especially thank Leonard Dung, Ian Robertson, and Christian Michel for their great advisory work.

We are also grateful to benefit from funding from the Alexander von Humboldt Foundation.

Thank you for joining us at the start of this project. We are excited to see where the conversation goes.

Guido Löhr

Co-Editor-in-Chief (together with Vincent Müller)

References

Babushkina, D. (2025). AI, decisions, and the reasons to believe: ethics-through-epistemology approach. *Philosophy of AI*, *1*(Philosophy of AI, Vol. 1 (2025)), 41–58. https://doi.org/10.18716/OJS/PHAI/2025.2276

Cappelen, H., Goldstein, S., & Hawthorne, J. (2025a). AI Survival Stories: a Taxonomic Analysis of AI Existential Risk. *Philosophy of AI*, *1*(1), 1–19. https://doi.org/10.18716/OJS/PHAI/2025.2801

Cappelen, H., Goldstein, S., & Hawthorne, J. (2025b). AI Survival Stories - Responses to Critics. *Philosophy of AI*, *1*(Philosophy of AI, Bd. 1 (2025)), 100–106. https://doi.org/10.18716/OJS/PHAI/2025.11987

Dung, L. (2025). Estimating the probability of AI existential catastrophe: Converging on the answer from opposite ends. *Philosophy of AI*, *1*(Philosophy of AI, Vol. 1 (2025)), 93–95. https://doi.org/10.18716/OJS/PHAI/2025.2854

Green, M. (2025). Large Language Models and the Varieties of Meaning. *Philosophy of AI*, *1*(Philosophy of AI, Bd. 1 (2025)), 34–40. https://doi.org/10.18716/OJS/PHAI/2025.11652

Koch, S. (2025a). Babbling stochastic parrots? A Kripkean argument for reference in large language

models. *Philosophy of AI*, *1*(Philosophy of AI, Bd. 1 (2025)), 19–33. https://doi.org/10.18716/OJS/PHAI/2025.2325

Koch, S. (2025b). What does it take to establish reference in LLMs? Kripke vs. Austin (Response to Green). *Philosophy of AI*, *1*(Philosophy of AI, Vol. 1 (2025)), 71–76. https://doi.org/10.18716/OJS/PHAI/2025.11963

Sauchelli, A. (2025). Artificial Intelligence, Personal Ontology, and Existential Risks. *Philosophy of AI*, *1*(Philosophy of AI, Vol. 1 (2025)), 59–70. https://doi.org/10.18716/OJS/PHAI/2025.2338

Sterri, A., & Skjelbred, P. (2025). Automation and its Discontents: Leveraging Automation to Safeguard Humanity. *Philosophy of AI*, *1*(Philosophy of AI, Bd. 1 (2025)), 86–92. https://doi.org/10.18716/OJS/PHAI/2025.11844

Svarc, R. (2025). Defending Alignment: A Commentary On 'AI Survival Stories.' *Philosophy of AI*, *1*(Philosophy of AI, Vol. 1 (2025)), 76–85. https://doi.org/10.18716/OJS/PHAI/2025.3327

Vredenburgh, K. (2025). AI survival stories, types of risk, and the precautionary principle. *Philosophy of AI*, *1*(Philosophy of AI, Vol. 1 (2025)), 96–99. https://doi.org/10.18716/OJS/PHAI/2025.11986