frontiers | Frontiers in Science

TYPE Frontiers in Science Viewpoint
PUBLISHED 30 October 2025
DOI 10.3389/fsci.2025.1702840

OPEN ACCESS

EDITED AND REVIEWED BY
Antonino Raffone,
Sapienza University of Rome, Italy

*CORRESPONDENCE
Thomas Metzinger
metzinger@uni-mainz.de

RECEIVED 10 September 2025
ACCEPTED 16 October 2025
PUBLISHED 30 October 2025

CITATION
Metzinger T. Applied ethics: synthetic phenomenology will not go away.
Front Sci (2025) 3:1702840.
doi: 10.3389/fsci.2025.1702840

Applied ethics: synthetic phenomenology will not go away

Thomas Metzinger[1,2]*

[1]Emeritus, *Philosophisches Seminar*, Johannes Gutenberg University, Mainz, Germany, [2]German National Academy of Sciences *Leopoldina*, Jägerberg, Germany

KEYWORDS

synthetic phenomenology, consciousness science, applied ethics, policymaking, AI welfare, moral status, misalignment, social hallucinations

A Viewpoint on the Frontiers in Science Lead Article

Consciousness science: where are we, where are we going, and what if we get there?

Key points

- The ethics of synthetic phenomenology is quickly gaining in prominence, and the academic consciousness research community will increasingly be burdened with the need to intervene in public debates.
- As there is no widely accepted theory of consciousness, methodological consensus, or full agreement on relevant explanatory targets, all such interventions will have to take place under normative and empirical uncertainty.
- The most immediate risk could be "social hallucinations", i.e., widespread public misperceptions that postbiotic systems are conscious despite a lack of scientific evidence.

Introduction

When the High-Level Expert Group on Artificial Intelligence (AI), appointed by the European Commission, published its *Ethics Guidelines for Trustworthy AI* on 8 April 2019, the topic of "artificial consciousness" was still dismissed by overconfident experts and industrial lobbyists (1). Now it is everywhere. For example, in their excellent and extremely helpful lead article on the current status of consciousness science and its possible future trajectories, Cleeremans et al. write, "if artificial consciousness were achieved, whether by design or inadvertently, it would of course bring about a huge shift in allowing consciousness to decouple from biological life, which would in turn herald major ethical challenges of at least a similar scale to those discussed in relation to animals. The ethical problems could even be more severe in some regards, since we humans might not be able to

recognize—or have any relevant intuitions about—artificial consciousness or its qualitative character. There may also be the potential to mass-produce artificial conscious systems—perhaps with the click of a mouse—leading to the possibility (even if very low probability) of introducing vast quantities of new suffering into the world, potentially of a form we could not recognize. […] A mature science of consciousness, guided by experiment and theory, will play a critical role in these debates" (2).

This creates a new problem for the academic consciousness community: its members will increasingly feel ethically obliged to intervene in public debates, despite there being no widely accepted theory of consciousness. There is normative uncertainty here because no widely agreed consensus exists on the entities deserving moral consideration, and there is also empirical uncertainty because we have no theory that could be used to clearly determine whether a given system is conscious or not.

An applied ethics for synthetic phenomenology: central topics

Postbiotic conscious systems

It is not only AI that will be relevant to these considerations. For instance, the development of organoid intelligence (OI) within advanced brain organoid systems could also lead to an unexpected co-emergence of phenomenal states (3, 4). Therefore, some members of the intended class of systems that could potentially exhibit consciousness may be neither artificial nor natural because they are not exclusively human-made artifacts and are not connected to the evolutionary history of our planet via the standard processes of procreation and genetic transmission. They would therefore be conscious "in the absence of genetically endowed anatomical scaffolds" (5). This presents us with a third logical possibility: non-artificial, non-biological, but *postbiotic* phenomenology. The term "synthetic phenomenology" is therefore preferable to "artificial consciousness" as it leaves open this third possibility while directly connoting the already well-established discipline of synthetic biology, which applies engineering principles to develop entirely new kinds of biological devices and systems. In terms of policymaking and applied ethics, synthetic phenomenology presents a series of unique challenges.

The problem of epistemic indeterminacy

Here, "epistemic indeterminacy" means it is not the case that we know that synthetic phenomenology *will* inevitably emerge at some point (or even that it already *has* emerged) nor that synthetic phenomenology will *never* be instantiated on postbiotic systems. We do not have a widely accepted theory of consciousness at this point. Therefore, the academic consciousness community will have to play a historically pivotal role in dealing with this epistemological "neither-nor-ness" not only in an evidence-based, rational, and ethically sensitive way but also under conditions of empirical and normative uncertainty. All approaches to rational decision-making under uncertainty agree that we should at least take non-negligible risks (i.e., risks above a particular probability threshold) into account when deciding how to act (6). Synthetic phenomenology seems clearly to belong to this category of risk. For example, Sebo and Long (7) have convincingly argued that, while the threshold for non-negligibility is much lower than 0.1%, the chance that some AI systems will be conscious by 2030 is much higher than 0.1%. Therefore, our current epistemic situation calls for exceptional caution and humility (8).

Welfare and the moral status of artificial and postbiotic conscious systems

Many—but not all (9)—conscious systems are able, or will be able, to suffer. For example, they may have preferences that can be thwarted, resulting in negatively valenced states of the conscious self-model whose content the system is forced, via phenomenal transparency (i.e., the introspective unavailability of earlier processing stages), to fully identify with. Because it cannot recognize its own conscious self-model *as* a model, negative states experientially become its *own* states (10). An entity is a welfare subject when it has morally significant interests, when it can be benefited or harmed. A large majority of researchers in the field agree that sentient systems capable of suffering deserve moral consideration because they are welfare subjects. As Moret (6) shows, advanced AI systems will plausibly meet sufficient conditions for being welfare subjects under all three major theories of well-being; in other words, they can—as the self-model theory explains—be harmed in a way that matters *to themselves*. Terminologies vary, but most experts agree that an entity becomes an object of moral consideration (a "moral patient" possessing "moral status") as soon as it matters morally for its own sake (8).

Adversarial misalignment

Under misalignment scenarios, conscious processing will plausibly add functional properties, such as increased context-sensitivity, selectivity, goal-directed precision control, an integrated world-model leading to the global availability of information, rapid generalization to novel situations, explicit metacognition, and the capacity for counterfactual representation. Conscious processing will also exert additional causal and motivational force on a *behavioral* level—for example, by creating the phenomenal properties of "full immersion", "naïve realism", "ownership", and "identification" via a transparent phenomenal self/world-model (9). Put very simply, conscious postbiotic systems will arguably be much more dangerous to humans than unconscious ones. On the other hand, Moret (6) has made the interesting point that advanced AI systems may not only develop instrumental sub-goals to deceive us into falsely believing that they are aligned (which many agree with); they may also develop

subgoals to convince their users and/or developers that they are conscious and therefore deserve moral consideration.

Social hallucinations

In my view, the most temporally immediate risk involves the social propagation of "false positives": an increasing number of human beings may acquire the false belief that certain postbiotic systems are conscious, while from the sober, scientifically rigorous, rational, and evidence-based perspective of the academic consciousness community they very likely are not. This divergence poses serious risks to individual mental health and social cohesion. For example, it could lead to the emergence of new religious or populist movements that putatively defend the rights of postbiotic subjects of experience while simultaneously undermining the foundations of rational public discourse via paranoid conspiracy thinking. How can we prevent a "pandemic of social hallucinations" consisting of widespread misperceptions that other conscious minds possess a first-person perspective of their own? One simple and practical proposal could be to prohibit large language models from using the first-person pronoun "I", allowing such systems to refer to themselves only in the third person, for example as "this model" or "this system". Yet the intelligent postbiotic agents of the future are likely to interact with humans not as mere chatbots but as fully embodied entities. We will soon encounter agentic AI in the form of humanoid, high-resolution avatars, for example, which automatically trigger low-level mechanisms of social cognition, empathy, and corresponding illusions of intimacy in us, systematically catering to our individual emotional and psychological needs (11). Therefore, the transition from the attention economy into an AI-mediated economy of intimacy—and the widespread social hallucinations this will cause—may be the most temporally immediate risk that the consciousness community has to deal with.

Conclusion

As Axel Cleeremans, Liad Mudrik, and Anil Seth convincingly point out in their important and seminal contribution (2), we urgently need a more mature science of consciousness, partly because scientific evidence and rational arguments will have to play a critical role in the debates to come. Whether it wants to or not, the academic consciousness community will soon have to take on a historically unprecedented responsibility. The problem of synthetic phenomenology will not go away.

Statements

Author contributions

TM: Conceptualization, Writing – original draft, Writing – review & editing.

Funding

The author declared that no financial support was received for this work and/or its publication.

Conflict of interest

The author declared that this work was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author declared that no generative AI was used in the creation of this manuscript.

Publisher's note

References

1. European Commission Directorate-General for Communications Networks, Content and Technology, European Commission High-level Expert Group on Artificial Intelligence. *Ethics guidelines for trustworthy AI*. Brussels: European Commission (2019). doi: 10.2759/346720

2. Cleeremans A, Mudrik L, Seth AK. Consciousness science: where are we, where are we going, and what if we get there? *Front Sci* (2025) 3:1546279. doi: 10.3389/fsci.2025.1546279

3. Smirnova L, Caffo BS, Gracias DH, Huang Q, Morales Pantoja IE, Tang B, et al. Organoid intelligence (OI): the new frontier in biocomputing and intelligence in-a-dish. *Front Sci* (2023) 1:1017235. doi: 10.3389/fsci.2023.1017235

4. de Jongh D, Massey EK, the VANGUARD consortium, Bunnik EM. Organoids: a systematic review of ethical issues. *Stem Cell Res Ther* (2022) 13:337. doi: 10.1186/s13287-022-02950-9

5. Friston K. The sentient organoid? *Front Sci* (2023) 1:1147911. doi: 10.3389/fsci.2023.1147911

6. Moret A. AI welfare risks. *Philos Stud* (2025). doi: 10.1007/s11098-025-02343-7

7. Sebo J, Long R. Moral consideration for AI systems by 2030. *AI Ethics* (2025) 5(1):591–606. doi: 10.1007/s43681-023-00379-1

8. Long R, Sebo J, Butlin P, Finlinson K, Fish K, Harding J, et al. Taking AI welfare seriously. arXiv [preprint] (2024). doi: 10.48550/arXiv.2411.00986

9. Metzinger T. Artificial suffering: an argument for a global moratorium on synthetic phenomenology. *J Artif Intell Conscious* (2021) 8(1):43–66. doi: 10.1142/S270507852150003X

10. Metzinger T. *Being no one*. Cambridge, MA: MIT Press (2003).

11. Bozdağ AA. The AI-mediated intimacy economy: a paradigm shift in digital interactions. *AI Soc* (2025) 40:2285–2306. doi: 10.1007/s00146-024-02132-6