Questioning the Moratorium on Synthetic Phenomenology

Roman Krzanowski[1]

1 Pontifical University of John Paul II in Kraków

Funding: No specific funding was received for this work.

Potential competing interests: No potential competing interests to declare.

Abstract

In a paper published in 2021, Prof. Thomas Metzinger proposed a global moratorium on synthetic phenomenology, beginning in 2021 and lasting until 2050, arguing that the development of conscious artifacts would bring about an explosion of human-like suffering on a global scale. We posit that Prof. Metzinger's argument for a global moratorium, as it has been formulated, is not justified. However, Prof. Metzinger's proposal, putting aside his claims about an explosion of suffering, can be taken as a generic warning against the uncontrolled development of AI - a position we fully agree with.

Keywords: synthetic consciousness, moratorium on synthetic phenomenology, synthetic suffering, category mistake, cognitive equivalence, Thomas Metzinger.

1. Introduction

Thomas Metzinger proposed a global moratorium on synthetic phenomenology beginning in 2021 and lasting until 2050 (Metzinger 2021a, b). We question whether Metzinger's call for a moratorium is justified and whether it can be regarded as a sound philosophical proposition.

Why is developing synthetic phenomenology so controversial that Metzinger believes it warrants a ban?[1] From an AI perspective, having a system with a sort of consciousness, which is defined as a set of cognitive-like capacities,[2] is an attractive proposition because it potentially means artificial systems with cognitive-like capacities would be more proficient at many human-like tasks than systems without these capacities (see e.g., Alexander 2020, Cali 2022). But Metzinger claims that developing synthetic conscious systems will unleash a flood of suffering for which we will be responsible and, therefore, we should not do it. How is that so?

Metzinger claims that, like any conscious entities, natural or artificial, which are cognitively equivalent to human (or biological) conscious agents (e.g., Shevlin 2020), these systems would be like us, or at least sort of like us, in that they would be thinking and feeling beings, as we are. Furthermore, these artificial systems would have phenomenological experiences (e.g., internal experiences, the concept of the self, self-knowledge, feelings, etc.) like we have. More specifically, as suffering is a common feature of conscious beings, at least of those we know about, these artificial systems would also have the capacity to suffer. Now, as "every conscious suffering being is a moral patient[3] demanding consideration" (Metzinger 2021a, b), these artificial creations would demand moral consideration from us as well, for at least two reasons: as suffering entities and as our creations (we may say our offspring) (e.g., Basl 20130). Thus, we will have a moral duty to these conscious artificial systems, whether they be disk drives, PCs, smartphones, WBE systems, bots, or whatever (e.g., Sotala and Gloor 2017), i.e., we are morally responsible for their wellbeing. Without knowing how to take care of suffering artificial systems, we should impose a moratorium, which is what Metzinger is asking for, on developing conscious artefacts with the innate capacity for suffering that comes with consciousness.

In the paper, we first summarize the major aspects of synthetic phenomenology relevant to the discussion. Next, we present Thomas Metzinger's argument for a global moratorium. We then consider whether the call for a global moratorium on synthetic phenomenology, as formulated by Thomas Metzinger, is a sound philosophical argument. Concluding, we propose adopting a charitable interpretation of Metzinger's memorandum

that would fit well with the current discussion about the dangers of AI and emphasize Metzinger's valuable contribution to the discussion on the uncontrolled development of AI technology and its impact on the humanities.

Metzinger's argument may be interpreted in two ways (at least). One way is to engage in the deep ethical dispute on the role of pain in all (natural and artificial) sentient beings and our moral duties towards suffering of any kind - existing, not existing, real, and imaginary. This is the way Metzinger wants us to face his argument. Another way to engage with Metzinger's argument, and this is our approach, is to resist being pulled into the deep moral grounds about abstract moral duties and the suffering of non-existent synthetic systems, and show that Metzinger's appeal for a global memorandum on synthetic consciousness is not based on factual analysis and therefore his argument does not warrant a deep ethical response, and in particular, it does not warrant industrial action, of which the ban on developing synthetic conscious systems is an example.

Two concepts - cognitive equivalence and category mistake - are critical for the discussion. By "cognitive equivalence," we mean a 100% equivalence between cognitive systems, not just functional equivalence or some other sort of qualified equivalence that excludes properties like feelings, emotions, free will, self-knowledge, and so on. The cognitive equivalence principle claims that any system with cognition will have the same cognitive functions, regardless of its origins, physical substrate, or creation history (see e.g., Shevlin 2020).

The second key concept is that of a category mistake or category error. The term is sought to originate with Ryle (Ryle 1942), but it has been long present in philosophy without having a specific label (see e.g., Boutler 2019, Magidor 2022). After Magidor (2016, 2022), we prefer to give examples of it rather than a definition. Magidor gives two: 'The number two is blue' and 'The theory of relativity is eating breakfast'. Such claims seem odd outside of poetry or literature. They strike us as incoherent claims on reality. Similarly, modern examples of this error are claims such as "Robots have feelings", "Robots are moral machines", "AI systems have free will", "Robots have rights", "Synthetic autonomous agents", or "Synthetic consciousness". Stevenson (Stevenson 2010) characterizes such claims as "The error of assigning to something a quality or action which can only properly be assigned to things of another category, for example, treating abstract concepts as though they had a physical location."

Category mistake is committed when there are alleged similarities between two concepts where there is none, as the compared concepts belong to different conceptual domains. For example, the claim that "Spring is happy" is a category mistake attributing human feelings to seasons. As an argument in analytical philosophical work about the nature of seasons, this claim would not fly. But as a line in poetry or uplifting writings, it would be quite well placed. Does it matter? It matters in studies where logic, accuracy, and practical actions matter[4].

Accepting claims based on category error without qualifications, though it may offer an interesting intellectual challenge, leads eventually to confusion, in particular if these claims spill over to claims about reality (see e.g., Velize 2012). Maybe the last quotation from Alexander will clarify this concept further: "The category mistake occurs when commentators equate conscious machines with living ones". And further, " … the error is due to the fact that the living makeup creates specific drives and needs during the process of becoming conscious in the living system, which is not the case with a designed machine" (Alexander 2020). Metzinger's claims equating suffering in natural and artificial systems, in our view, are such a category error, and on this category error, Metzinger is basing his argument. For a more detailed discussion of category error or category mistake, see Blackburn (1994), Honderich (2005), Audi (2015), and Magidor (2016, 2022).

2. What Is Synthetic Phenomenology?

What is synthetic phenomenology (SP)? There are many definitions of SP, as one may well expect with philosophical and psychological concepts that have been adopted in cross-disciplinary research fields like AI. With some approximation, the definitions of SP generally fall into one of two categories:

I. Those denoting it as a research program in AI and robotics (Gamez 2008, Chrisley and Parthemore 2007, Alexander 2020, Cali 2022).
II. Those denoting it as an abstract concept that describes the phenomenal properties of any conscious system, whatever it may be, such as biological, mechanical, or something else (e.g., post-biotic) (Dennett 1991, Alexander and Morton 2007, Christley 2009, Chalmers 2017, Metzinger 2021, Smith and Schillaci 2021).

No classification is ever perfect, so some definitions may blend, to varying degrees, elements of both these categories[5]. Thus, allocating them to (I)

or (II) would depend on how the employed terminology is interpreted.

The first definition (I) is quite clear about what it is pursuing, namely consciousness-like functions in artefacts rather than conscious first-person experiences. The latter definition (II) generalizes consciousness to any object or system, thus implicitly assuming that whatever the carrier of the consciousness is (i.e., human agent, artefact, silicon, or post-biotic system), its consciousness will have the same capacities as the original concept, so it assumes that SP is multi-realizable. In other words, research on one kind of conscious system would give insights into all others, and modelling a consciousness in one system will also provide models for other systems. Metzinger's synthetic phenomenology falls into the second (II) category of SP, and this seems to be the reason for his argument.

3. Metzinger's Argument

Thomas Metzinger proposed a global moratorium on synthetic phenomenology from 2021 to 2050 (Metzinger 2021a, b). His argument goes as follows (Metzinger 2021a):[6]

> *(A1) On ethical grounds, we should not risk a second explosion of conscious suffering on this planet, at least not until we have a much deeper scientific and philosophical understanding of what both* ***consciousness*** *and* ***suffering*** *really are.*
>
> *(A2) As we presently have no* ***good theory of consciousness*** *and no good, hardware-independent* ***theory about what suffering*** *really is, the risk of ENP (explosion of negative phenomenology) is currently incalculable.*
>
> *(A3) It is unethical to incur incalculable (of explosion of negative phenomenology)) risks of this magnitude.*
>
> *Therefore,*
>
> *(C1) Until 2050, there should be a global ban on all research that directly or indirectly aims, or knowingly risks, the emergence of synthetic phenomenology (as synthetic phenomenology would bring an explosion of conscious suffering).*[7]

The central concept behind Metzinger's argument for a moratorium is the suffering of conscious beings. According to him, with his type II definition of SP, any conscious system suffers.[8] So synthetic systems or conscious artefacts of any sort, conscious silicon things, or conscious post-biotic systems, once created, will also suffer as conscious human agents do. We have no idea how synthetic systems would suffer, or whether they would suffer at all, but at least according to Metzinger and his cognitive equivalence principle, we cannot exclude the possibility of suffering on logical or technical grounds. Thus, Metzinger adopts the precautionary principle in assuming that synthetic systems or artefacts with synthetic consciousness will suffer, in the same way (employing the cognitive equivalence principle!) as natural conscious systems do.

For suffering to occur in any kind of system, four necessary conditions must be satisfied (Metzinger 2021a): (I) The C condition: a conscious experience,[9] (II) The PSM condition: possession of a phenomenal self-model (PSM),[10] (III) The NV condition: negative valence,[11] and (IV) The T condition: transparency.[12]

To these four conditions, Metzinger adds four corollaries:

a. An unconscious system is unable to suffer.
b. A conscious system without a coherent PSM is unable to suffer.
c. A self-conscious system without the ability to produce negatively valenced states is unable to suffer.
d. A conscious system without any transparent phenomenal states cannot suffer, because it would lack the phenomenology of ownership and identification (Metzinger 2021a).

Therefore, before any system can experience suffering, it must satisfy four conditions (I–IV) and four corollaries (a–d). Clearly, humans do satisfy these, as well as some animals. But Metzinger claims that any system with consciousness, synthetic or natural, would also satisfy them, regardless of its origin, creation history, or substrate, whether it be biological, silicon, some hybrid, or some hitherto unknown substrate.

Metzinger states that any conscious (with natural or synthetic consciousness) system will suffer[13]. It is then logical (meaning that there are no logical contradictions) to assume that artificial conscious systems, of which we know nothing, and that do not exist yet but may one day exist and be conscious, may suffer as well. Thus, any systems with consciousness, synthetic or not, existing now or in the future, will suffer (as human agents do) and, when these synthetic systems populate the Earth, we will witness an explosion of suffering on Earth (i.e., an explosion of negative phenomenology realized in artificial conscious systems).

As moral agents, Metzinger posits further, we have duties towards any creatures and our creations (i.e., artefacts) in particular, and we should strive to minimize suffering in general and the suffering of these artificial creatures. Thus, we should stop working on developing conscious artefacts to avoid an alleged explosion of negative phenomenology in synthetic systems of our creation. We should do it at least until we understand what suffering is, what these artificial systems really are, and whether (and how) they suffer. Metzinger's argument is much more elaborate than this short exposition. But this summary offers the gist of it.

4. Argument Dissected

Let us dissect Metzinger's argument and emphasize its pivotal claims, as we see them.

(Claim a) Matzinger begins with an assumption (implicit) that human agents suffer and that natural (not hybrid, or chimeras) biological conscious systems suffer as well (Metzinger 2017).

This claim is more or less (we are not sure how animals suffer) based on our experience. We can accept the claim as factual with the qualification that consciousness (and everything that comes with it) is restricted to a certain (rather vaguely defined) level of biological organization (excluding simple organisms)[14]

(Claim b) Then, Metzinger abstracts the concept of consciousness from its biological or natural foundations, claiming that consciousness is a functional feature, implementation independent, and can be present in any system with a set of certain properties.[15] Further, any conscious system is conscious the same way as natural biological conscious systems are, in all dimensions, including the capacity to suffer. This is the so-called cognitive equivalence (explained in previous sections).

(Claim c) Concluding, as we are responsible for the wellbeing of our creations, we should stop working on synthetic consciousness to avoid the creation of these conscious objects until we better understand the problem of suffering in artefacts (artificially created objects)[16].

We face, in claims (a) to (c), factual (related to physical reality) and speculative (related to conceptual constructs) assumptions mixed. How is that so? The first claim (Claim a) is that conscious biological systems (as qualified in ft 13) clearly suffer. We know it as a fact because we all have direct experience of it. The second claim is that there is an epistemic possibility[17] (Derose 1991) that artificial conscious systems, of which we know nothing and can only assume may one day exist, may suffer as well (Claims b, c), as we do. But this claim is speculation (as opposed to a factual claim) based on two assumptions:

1. an assumption of conceptual similarities between natural and artificial consciousness (see the definition of cognitive equivalence in the introduction), and
2. an assumption that conceptual similarities (between natural and artificial consciousness) translate into factual similarities between natural systems and artifacts.

Let us look at these assumptions in detail:

(1) The assumption of cognitive equivalence of natural and artificial systems states that consciousness in natural systems is fully equivalent to that in artifacts. But without detailed, factual justification, such a claim is a factual error. We are talking here about two different realms: that of nature and that of artifacts (see Alexander 2020). In the assumption of cognitive equivalence, we blur the differences between two different categories of objects: conceptual entities and real objects.

Is the blending of different categories (conceptual and real) of objects permissible in an argument? Yes, as long as it is acknowledged and recognized. Is such a step productive? Yes, why not? In philosophical speculations, we often do it. We should not do it, and in general, we try to avoid it if we plan to make practical use of it or stake factual claims, as Metzinger does. There is no necessary (i.e., logical or epistemic necessity) relation between artificial consciousness and the consciousness of natural systems, and Metzinger admits this. Despite this, he still passes over the differences between humans and machines (by claiming cognitive equivalence), arguing from the nonexistent, speculative assertion (that synthetic conscious artifacts will suffer because biological ones do) to realities or facts, i.e., conscious artefacts. Some people accept this logic as valid and sound; we do not. We have posited that cognitive equivalence, while being an interesting concept, does not map into equivalent properties of natural and non-existing artifacts, nor is it confirmed by experimental observations. It is a conjecture. And again, as we said, as a speculative claim, such a conjecture is acceptable. But as a claim about reality, it is a factual error. Two different ontological realms (virtual and real) do not mix by fiat or argument[18].

While in many contexts conceptual discussions may be fruitful, in claims on reality asking for real action, pure conceptual claims will not do. This incongruence between claims and reality is quite widespread in the ongoing debate about the ethics of AI or the human-compatibility of AI systems (Dreyfus and Dreyfus 1988, Russell 2019, Veliz 2021, Klein 2023, Mickunas and Pilotta 2023). Such claims, unfortunately, are quite common in public fora, blogs, papers, the popular press, and even scientific publications related to AI technology, and they seem to obfuscate the real problems facing artificial intelligence (Russell 2019, Zuboff 2019, Veliz 2021, Powers and Ganascia 2020, Klein 2023, Mickunas and Pilotta 2023). But being common, popular, or widespread opinion does not make it right. If taken at face value, Metzinger's claims may have the same obfuscating effect on rational discussion about AI as the discussions about moral robots, free will in robotic systems, or feelings and emotions in synthetic creations.

There is nothing wrong with speculations like Metzinger's (they are the bread and butter of the philosophical enterprise) as long as they remain in the realm of philosophical meditation without making any substantial claims on reality[19]. But this is not the case with Metzinger; his claims are about reality, about real consequences – it is a call to action, and all this on a planetary scale.

5. Conclusions

There is nothing illogical or incoherent about synthetic phenomenology when it is defined as a research program within AI aimed at studying and developing phenomenological-like functionality in artificial systems (e.g., Gamez 2008, Chrisley 2009, Mlsbt 2021, Cali 2022). Indeed, synthetic phenomenology seems to be a farfetched, futuristic AI endeavor, like many other AI projects or even Turing's proposal of a thinking machine and McCarthy's AI project itself (McCarthy 1959). As well, there is nothing wrong about conceptual speculations per se. Conceptual mediations are quite ubiquitous in philosophy.

But Metzinger wants us to take his argument beyond mere speculations, as an ethical and factual problem, while we claim that there is none, or at the minimum, Metzinger does not show us that there is one, and the apparent gravity of the argument is a result of a sort of philosophical conjuring act – we create reality out of speculations. The error that Metzinger makes is that of claiming that the (assumed) conceptual similarity (of the synthetic phenomenology of artifacts and natural phenomenology) implies reality (with the phenomenology of real agents). We posit that this is a connection that is not justified by Metzinger, that he commits a conceptual error, and his claim cannot be taken or accepted as stated by him.

The paper attempts to show that the apparent conceptual identity does not legitimize identity in reality.[20] It is quite common in modern philosophy to brush over this distinction (between the conceptual and the real) and to accept, without much ground to do so, conceptual claims as legitimate claims on reality, as long as these claims do not violate logic; i.e., what is logically conceivable is conceivably real (see examples discussed by Bulter (2019)). These may be claims of similarity or difference[21].

One may suggest that accepting Metzinger's claim of artefacts suffering is the safe bet in the presence of epistemic ignorance, or as he calls it, epistemic indeterminacy. Yes, there is some truth to it, and we do not deny it. But epistemic indeterminacy characterizes, to varying degrees, any advanced AI research field, such as ethical AI, AGI, Explanatory AI (XAI), trusted AI, human-compatible AI, and provably beneficial AI (PBAI), and nobody calls to stop this work, even if these research programs are potentially threatening our very existence. Epistemic ignorance (as it always has

been) should be rather a stimulus for research, thought made carefully, than a ban.

In summary, we propose to look at Thomas Metzinger's claim not as a well-formed, sound philosophical argument but rather as an ideological manifesto warning about the potential dangers of the uncontrolled development of AI technology in general, especially if it is left in the hands of technologists, businesses, military establishments, and moneyed interests. Under such an interpretation, Thomas Metzinger could join the growing list of prominent researchers, thinkers, and sci-fi writers voicing concerns about the dangers of an unmitigated development and deployment of AI systems, even if these systems are not AGI-like yet (e.g., Lem 2014/1965, Russell 2019, Gupta 2020, Bartoletti 2021, Gawdat 2023, Hinton et al. 2023). This is maybe where the significant, undisputable merit of his work lies. But a global moratorium on synthetic phenomenology is not justified, desired, or productive.

Acknowledgements

The author would like to thank the anonymous reviewer who pointed out some flaws in the argument in the first version of the paper. In response, the authors rewrote the discussion, as well as deleted some sections that took the reader off the main story. All of these changes, we hope, added to the clarity of the paper. But it is the readers who will judge, not the author.

Footnotes

[1] For the purposes of this discussion, synthetic phenomenology (SP) is synonymous with artificial consciousness. "The terms 'phenomenology' and 'consciousness' suggest that the capability for having experiences, namely of having states whose phenomenal content enables the artificial agent to have access from its point of view to the surrounding world, is crucial to move, make decisions, and carry out actions effectively" (Cali 2022).

[2] We use terms like "sort of" and "-like" to indicate that the qualified term is not being used in its original meaning but rather as a vague extension. Thus, for example, "a sort of consciousness" refers to a concept of consciousness that resembles the original concept but is not exactly equivalent. This "softening" of meaning is a standard philosophical practice among some philosophers and AI engineers, in particular when working with problems that span multiple fields of research (e.g., Chalmers 2017, Dennett 2018).

[3] To be a moral patient is to be "the target of the actions of a moral agent, and [to] be worthy of moral consideration" (Floridi and Sanders 2004).

[4] Revisit the discussion of the category mistake in the introduction.

[5] One may recall similar conceptualizations in artificial intelligence with weak, strong AI (see Searle 1984, 1990) and in-between AI for AI systems that escape the rigid two-class structure (which most of modern AI systems do).

[6] All quotations, if not otherwise stated, are from Metzinger's paper on Artificial Suffering (Metzinger 2021a).

[7] By the first explosion of conscious suffering, Metzinger refers to the emergence on Earth of biological conscious organisms with the inherent capacity to suffer. A second explosion of conscious suffering would happen if we were to create conscious systems that, according to the cognitive equivalence principle, would have the capacity to suffer. (Note that the concept of suffering used in Metzinger's argument is an abstraction borrowed from biological systems like humans.) The second explosion of negative phenomenology refers to a "second explosion of conscious suffering on this Planet" in "advanced AI and other post-biotic systems" (Metzinger 2021a).

[8] Metzinger is not alone in attributing the ability to suffer to synthetic systems. For example, David Chalmers claimed that conscious machines (with synthetic consciousness) will feel pain the same way we do, happiness the way we do, etc. (Chalmers 2017, min 2:43). We need to add that this will be true if and only if synthetic consciousness is equivalent to human consciousness, about which we know some things, or animal consciousness, about which we know somewhat less, because these are the only forms of consciousness that we know suffer in some way, so we can only refer to these.

[9] "Suffering" is a phenomenological concept, and only beings with conscious experience and a phenomenal self-model (PSM) can suffer. Zombies

do not suffer, and human beings in a dreamless deep sleep, in a coma, or under general anesthesia do not suffer. It is also possible that persons or unborn human beings who have yet to come into self-conscious existence do not suffer. Robots, AI systems, and intelligent post-biotic entities can only suffer if they are capable of having phenomenal states (Metzinger 2021a).

[10] "The most important phenomenological characteristic of suffering is the sense of ownership, the untranscendable subjective experience that it is myself who is suffering right now, that it is my own suffering I am currently undergoing" (Metzinger 2021a).

[11] "The suffering system must be able to internalize and integrate the negative value of an experience." In other words, "Suffering is created by states representing a negative value being integrated into the PSM of a given system" (Metzinger 2021a).

[12] "Phenomenal transparency means that something particular is not accessible for subjective experience, namely the representational character of the contents of conscious experience" (Metzinger 2021a).

[13] Metzinger implicitly extends human experience of suffering to other forms of consciousness as this is the only experience of suffering that we directly have, and we can make factual claims about it.

[14] We accept that some will claim that all biological life has some level of consciousness, but we do not want to argue this and limit our concept of consciousness to higher order animals, knowing that this classification is not very crisp. Similarly, we also do not engage with the claims that nonliving objects, machines, computers, or other artifacts are conscious.

[15] See the four necessary conditions for suffering to occur in any kind of system (Metzinger 2021a).

[16] "But every entity that is capable of suffering should be an object of moral consideration…. We are ethically responsible for the consequences of our actions" (Metzinger 2021a).

[17] On epistemic modality, see Brandon (2023).

[18] Philosopher David Chalmers (2023) claims that the virtual is real, but his real is not physically real.

[19] As we may recall, Descartes's argument about mind-body separation was quite cogent (logically admissible) as long as Descartes did not try to translate it into the realities of the human body.

[20] Reality is what is the final arbiter of conceptual constructs about reality, as one can learn from conflicting theories about quantum mechanics (Becker 2019).

[21] As an example, we recall that the difference between the Morning Star and the Evening Star is only in concept, not in reality, and the alleged similarity of ethics of AI systems does not correspond to ethics in human agents without detailed qualifications and thorough analysis (see e.g., Veliz 2021).

References

- Aleksander, I. 2022. "From Turing to Conscious Machines" Philosophies 7, no. 3: 57 https://doi.org/10.3390/philosophies7030057
- Aleksander, I. 2020. The category of machines that become conscious, J. Artif. Intell. Conscious. 7(1), 313.
- Aleksander, I., and Morton, H. 2007. Why axiomatic models of being conscious? Journal of Consciousness Studies, 14, 15–27.
- Arabales, R., A. Redezma, and A. Sanchis. 2009. Establishing a roadmap and metric for conscious machine development. Published in: Proceedings of the 8th IEEE International Confenrence on Cognitive Informatics, Kowloon, Hong Kong, 15-17 June 2009, pp.94-101. https://e-archivo.uc3m.es/bitstream/handle/10016/10430/establishing_arrabales_ICCI_2009_ps.pdf;jsessionid=CFD777964ED614DF35B3E605F4C9F9DE?sequence=2
- Audi, R. 2015. The Cambridge Dictionary of Philosophy. 3rd.ed. Cambridge: Cambridge University Press.
- Bain, D., M. Brady and J. Corns. 2020. Philosophy of Suffering. Metaphysics, Value, and Normativity. London: Routledge.

- Bartoletti, I. 2021. An Artificial Revolution. London: The Indigo Press.
- Basl, J. 2013. The Ethics of Creating Artificial Consciousness. APA Newsletter on Philosophy and Computers 13 (1):23-29.
- Becker, A. 2019. What is Real? London: John Murray (Publishers).
- Bennet, M., D. Dennett, P. Hacker, and J. Searle. 2007. Neuroscience and Philosophy. New York: Columbia University Press.
- Blackburn, S. 1994. The Oxford Dictionary of Philosophy. Oxford University Press. p. 58.
- Bostrom, N. 2012. The superintelligent will: Motivation and instrumental rationality in advanced artificial agents. Minds and Machines, 22(2 - special issue 'Philosophy of AI' ed. Vincent C. Müller), 71–85.
- Brandon, C. 2023. Epistemic Modality. IEP. Available athttps://iep.utm.edu/ep-moda/
- Bulter, S. 2019. Why Medieval philosophy matters? London: Bloomsbury Publishers.
- Cali, C. 2022. Philosophical, Experimental and Synthetic Phenomenology: The Study of Perception for Biological, Artificial Agents and Environments. Foundations of science. https://doi.org/10.1007/s10699-022-09869-7
- Chalmers, D. 2017. Artificial Consciousness — David Chalmers. Available athttps://www.youtube.com/watch?v=RIAIuv31YKs
- Chalmers, D. 2023. SuperIntelligence. Available at https://www.youtube.com/watch?v=hPQJUP52V4A
- Chrisley, R. 2009. Synthetic Phenomenology. International Journal of Machine Consciousness 2009 01:01, 53-70. DOI: 10.1142/S1793843009000074.
- Chrisley, R., & Parthemore, J. 2007. Synthetic phenomenology: Exploiting embodiment to specify the nonconceptual content of visual experience. Journal of Consciousness Studies, 14, 44–58.
- Davies, J. 2012. The Importance of Suffering: the value and meaning of emotional discontent. London: Routledge ISBN 0-415-66780-1
- Defense. 2002. "Defense.gov News Transcript: DoD News Briefing – Secretary Rumsfeld and Gen. Myers, United States Department of Defense (defense.gov)". February 12, 2002. Available at https://archive.ph/20180320091111/ http://archive.defense.gov/Transcripts/Transcript.aspx?TranscriptID=2636
- Dennett, D. 1991. Consciousness explained. Boston: Back Bay Books.
- Dennett, D. 2018. From Bacteria to Bach. London: Penguin.
- DeRose, K. 1991. Epistemic Possibilities. The Philosophical Review, 100(4), 581–605.https://doi.org/10.2307/2185175
- Dreyfus, H. L., and S. E. Dreyfus. 1988, Making a mind versus modelling the brain. In Drefyus, H. L. Skillful Coping. Oxford: Oxford University Press. pp. 205-230.
- Floridi L. and J.W. Sanders. 2004. On the Morality of Artificial Agents. Minds and Machines 14: 349–379.
- Frances, B. 2021. The Problem of Suffering. In: An Agnostic Defends God. Palgrave Macmillan, Cham.https://doi.org/10.1007/978-3-030-73331-5_6
- Gamez, D. 2008. Progress in machine consciousness, Consciousness and Cognition, Volume 17, Issue 3, 2008, Pages 887-910, ISSN 1053-8100, https://doi.org/10.1016/j.concog.2007.04.005
- Gawdat, M. 2013. Don't bring children into this AI world. EMERGENCY EPISODE: Ex-Google Officer Finally Speaks Out On The Dangers Of AI! - Mo Gawdat. Available at https://www.youtube.com/watch?v=bk-nQ7HF6k4
- Gupta, V. 2020. The Future Stuff. London: Unbound.
- Hinton, G. 2023. Statement on AI Risk. Open Letter. Available athttps://www.safe.ai/statement-on-ai-risk#open-letter
- Honderich, T. 2005. The Oxford Companion to Philosophy. 2nd ed. Oxford; Oxford University Press.
- Hopkins, P. D. 2012. Why uploading will not work, or, the ghosts haunting transhumanism. International Journal of Machine Consciousness. Vol. 4, No. 1 (2012) 1250014.
- Klein, N. 2023. AI machines aren't 'hallucinating'. But their makers are. The Guardian. Available at https://www.theguardian.com/commentisfree/2023/may/08/ai-machines-hallucinating-naomi-klein
- Kleiner, J. 2020. "Mathematical Models of Consciousness" Entropy 22, no. 6: 609.https://doi.org/10.3390/e22060609
- Koene, R.A. 2012. "Fundamentals of Whole Brain Emulation: State, Transition and Update Representations". International Journal on Machine Consciousness Vol. 4, No. 1 (2012).pp 5-21.
- Koene, R.A. 2013. Uploading to Substrate-Independent Minds. The Transhumanist Reader: Classical and Contemporary Essays on the Science,

Technology, and Philosophy of the Human Future, First Edition. Edited by Max More and Natasha Vita-More. John Wiley & Sons, Inc. pp. 146-156.

- Krzanowski, R. and P. Polak.2023. Philosophy in Technology: Objectives, Questions, Methods, and Issues. Workshop on Philosophy in Technology: The Philosophical Challenges for Technology from Various Points of View, April 28–29, 2023. Wrocław University of Science and Technology. Available at https://www.researchgate.net/publication/370653723_Philosophy_in_Technology_Objectives_Questions_Methods_and_Issues
- Langle, A. 2008. Suffering—an Existential Challenge: Understanding, dealing and coping with suffering from an existential-analytic perspective. International Journal of Existential Psychology & Psychotherapy. Volume 2, Issue 1. Available at https://www.meaning.ca/web/wp-content/uploads/2008/01/115-13-486-1-10-20171212.pdf
- Lem, S. 2014/1965. The Cyberiad. London: Penguin Books.
- Lewis, C.S. 2001/1940.The Problem of Pain. San Francisco: Harper.
- Magidor, O. 2016. Category Mistakes. Oxford: Oxford University Press.
- Magidor, O. 2022. "Category Mistakes", The Stanford Encyclopedia of Philosophy (Fall 2022 Edition), Edward N. Zalta & Uri Nodelman (eds.), URL = < https://plato.stanford.edu/archives/fall2022/entries/category-mistakes/>
- Marcus, G. 2022. Artificial General Intelligence Is Not as Imminent as You Might Think. Scientific American. Available at https://www.scientificamerican.com/article/artificial-general-intelligence-is-not-as-imminent-as-you-might-think1/
- McCarthy, J. 1959. "Programs with Common Sense" at the Wayback Machine (archived October 4, 2013). In Proceedings of the Teddington Conference on the Mechanization of Thought Processes, 756–91. London: Her Majesty's Stationery Office.
- Metzinger T. 2008. Empirical perspectives from the self-model theory of subjectivity: a brief summary with examples. Prog Brain Res. 2008;168:215-45. doi: 10.1016/S0079-6123(07)68018-2. PMID: 18166398.
- Metzinger T. 2003. Being No One. The Self-Model Theory of Subjectivity. Cambridge: MIT Press.
- Metzinger T. 2007. Self models. Scholarpedia, 2(10):4174. Available athttp://www.scholarpedia.org/article/Self_models
- Metzinger T. 2017. Suffering.In Kurt Almqvist & Anders Haag (2017)[eds.], The Return of Consciousness. Stockholm: Axel and Margaret Ax:son Johnson Foundation. ISBN 978-91-89672-90-1
- Metzinger, T. 2021. Why we should worry about computer suffering. IAI News. /articles/why-we-should-worry-about-computer-suffering-auid-1761.
- Metzinger, T. 2021a. Artificial Suffering: An Argument for a Global Moratorium on Synthetic Phenomenology. (Philosophisches Seminar, Johannes Gutenberg-Universität Mainz, D-55099 Mainz, Germany) Journal of Artificial Intelligence and Consciousness 2021 08:01, 43-66. https://doi.org/10.1142/S270507852150003X
- Metzinger, T. 2021b. Three Types Of Arguments for a Global Moratorium on Synthetic Phenomenology. Pufendorf lecture at the Department of Philosophy, Lund University, 21 October 2021. Available at https://www.youtube.com/watch?v=RzhpmAlMURQ
- Mickunas, A. and J. Pilott. 2023. A Critical Understanding of Artificial Intelligence: A Phenomenological Foundation. Singapore: Bentham Science Publishers Pte. Ltd.
- Mlsbt. 2021. The problem of artificial suffering. Effective Altruism Forum. Available at https://forum.effectivealtruism.org/posts/JCBPexSaGCfLtq3DP/the-problem-of-artificial-suffering
- Müller, V. C., & Cannon, M. 2022. Existential risk from AI and orthogonality: Can we have it both ways? Ratio, 35, 25– 36. https://doi.org/10.1111/rati.12320
- Powers, T. M., and Jean-Gabriel Ganascia, 'The Ethics of the Ethics of AI', in Markus D. Dubber, Frank Pasquale, and Sunit Das (eds), The Oxford Handbook of Ethics of AI (2020; online edn, Oxford Academic, 9 July 2020), https://doi.org/10.1093/oxfordhb/9780190067397.013.2, accessed 13 May 2023
- Russell, S. 2019. Human Compatible. AI and problems of control. London: Penguin.
- Ryle, G. 1942. The Concept of mind. Routledge edition (2009). New York: Routledge.
- Sager, A. R. 2021. "The Existential Problem of Evil: Theodicy, Theosis, and the Threat of Meaninglessness" (2021). ETD Collection for Fordham University. AAI28496133. Available at https://research.library.fordham.edu/dissertations/AAI28496133
- Sandberg, A. 2013. Feasibility of Whole Brain Emulation. Philosophy and Theory of Artificial Intelligence, 251–264. doi:10.1007/978-3-642-

31674-6_19.

- Sandberg, A. and N. Bostrom. 2008. Whole Brain Emulation: A Roadmap, Technical Report #2008-3, Future of Humanity Institute, Oxford University. Available at www.fhi.ox.ac.uk/reports/2008-3.pdf.
- Schneider, S. 2020. How to Catch an AI Zombie In: Ethics of Artificial Intelligence. Edited by: S. Matthew Liao, Oxford University Press (2020). © Oxford University Press. DOI: 10.1093/oso/9780190905040.003.0016
- Searle, J. R. 1984. Minds Brains, and Science, Penguin, London.
- Searle, J. R. 1990. 'Is The Brain A Digital Computer?', Proceedings and Addresses of the American Philosophical Association 64(3), 21-37.
- Shevlin, H. 2019.To build conscious machines, focus on general intelligence: A framework for
- the assessment of consciousness in biological and arti¯cial systems," in Proc. Towards
- Conscious AI Systems Symposium, CEUR Workshop Proceedings, Vol. 2287, Paper 10
- (Palo Alto, CA), 8 pages.
- Smith D.H. and G. Schillaci. 2021. Why Build a Robot With Artificial Consciousness? How to Begin? A Cross-Disciplinary Dialogue on the Design and Implementation of a Synthetic Model of Consciousness. Front. Psychol. 12:530560. doi: 10.3389/fpsyg.2021.530560
- Sotala, K., and L. Gloor. 2017. Superintelligence as a Cause or Cure for Risks of Astronomical Suffering. Informatica 41 (2017) 389–400 389.
- Stevenson, A. (ed.). 2010. Oxford Dictionary of English, third edition, Oxford: Oxford University Press.
- Suffering n.d. What did the Buddha mean by suffering? Available athttps://tricycle.org/beginners/buddhism/what-did-the-buddha-mean-by-suffering/
- Tomasik, B. 2019. What are suffering subroutines? Avaiable athttps://reducing-suffering.org/what-are-suffering-subroutines/
- Veliz, C. 2021. Moral zombies: why algorithms are not moral agents. AI & Society (2021) 36:487–497https://doi.org/10.1007/s00146-021-01189-x
- Wasson. D. 2018. Roman Daily Life. Available athttps://www.worldhistory.org/article/637/roman-daily-life/
- Woodridge, A. 2020. The Road to Conscious Machines. London: Penguin.
- Zuboff, S. 2019. The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power (Campus, 2018; PublicAffairs, 2019).