Open Peer Review on Qeios

Philosophy in Technology: Objectives, Questions, Methods, and Issues

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

Philosophy in technology is a research program aimed at studying the philosophical roots of engineering and technology, and it asserts that the resolutions to problems need to be based on an understanding of these philosophical roots. In this paper, we define the objectives of philosophy in technology, the kinds of questions it seeks to explore, the methods it applies, and the ways in which it differs from the philosophy of technology. We then examine six selected problems to illustrate how the philosophical perspective can shine new light onto technology.

Keywords: philosophy in technology, philosophy of technology, semantic gap, synthetic phenomenology, orthogonality thesis, Searle's Chinese Room, meta-ontology of AGI systems.

1. Introduction

As a research program, philosophy in technology is concerned with the philosophical roots of engineering and technology. It is not specific to any particular technical domain, with it being more concerned about how different technologies can benefit from purely philosophical concepts, how technological domains often unwittingly adapt traditional philosophical concepts to meet their needs, and how from an abstract metaphysical, ontological, or axiological perspective, philosophy shapes and defines what technology does, how it develops, and how it evolves.[1]

Philosophy in technology also highlights the semantic gap that exists between the concepts used by technology and the concepts understood in philosophy. Our claim here is that this semantic gap is a source of confusion and misunderstanding between philosophers, the general population, and technologists. Moreover, it also serves to overinflate or underestimate the risks and threats posed by technological developments.

We begin in the following section by contrasting philosophy in technology with the philosophy of technology. Section 3 then discusses the main tenets of philosophy in technology as a research program, before Section 4 outlines the methodological assumptions of philosophy in technology. Section 5 then presents some exemplary cases of philosophical thinking in technology. Finally, Section 6 summarizes our observations of philosophy in technology and suggests a need for an open dialogue between philosophers and technologists, although they are actually not as far apart as many seem to think.

2. Philosophy in technology versus the philosophy of technology

The philosophy of technology can be viewed from many perspectives. For example, it can be seen as (1) a systematic clarification of the nature of technology as an element and product of human culture. Alternatively, it can be regarded as (2) a systematic reflection on technology's consequences for human life or (3) a systematic investigation of the practices involved in inventing, designing, engineering, and producing technological artifacts.

In contrast, philosophy in technology (1) seeks the implicit philosophical ground of technology and engineering together with the role it plays in shaping technological solutions; (2) explicates the ontological, metaphysical, axiological, and methodological dimensions of technology; and (3)

clarifies the semantic gap between technical and philosophical concepts in an attempt to bring them together under one perspective. The latter ambition could involve concepts such as agents, autonomy, intelligence, the mind, ethics, justification, responsibility, phenomenology, selfhood, personhood, knowledge, wisdom, privacy, power, right vs. wrong, ontology, truth conditions, verification, and so on, with this list being virtually endless.

3. Philosophy in technology as a research program

Philosophy in technology attempts to clarify the philosophical roots of technology by (A) explaining how philosophy is present in technology and engineering; (B) clarifying the role that philosophy plays in technology and engineering (i.e., philosophy for technology and engineering); (C) sparking a discussion about the philosophical foundations and implications of new technologies, such as for reducing any existential threats; (D) using philosophical reflection to shape technology that is more humanistic; (E) seeking to change the way philosophy is taught in technical universities; and (F) opening up the technical perspective to philosophical analysis.

4. Philosophy in technology: methodology

As a research program, philosophy in technology was created as an adaptation and extension of Michael Heller's concept of *philosophy in science*, which Heller developed in the 1980s to primarily analyze the relationship between philosophy and physics. Since then, this concept has proven to be very useful for highlighting the relevance of philosophy to other hard sciences. Nevertheless, on reflecting on the problems of modern technology, we saw that an analogous concept was needed to analyze the relationship between philosophy and technology.

But what is philosophy in technology? (A) Philosophy in technology (PinT) is a **reflection on classical concepts in technology**. It is analogous to philosophy in science because it proposes tracing the presence and role of the big classical philosophical questions in technology, such as the nature of free will, the mind, autonomous agents, and so on, so we can identify and analyze references to classical philosophical concepts like matter and time. Furthermore, philosophy in technology seeks to explore how classical philosophical concepts can be adapted to meet the needs of technology. (For example, Aristotelian phronetic ethics have been adapted for machine ethics, and the utilitarian and deontic ethical schools have been used in AI.) (B) Philosophy in technology represents a **disclosure and critical analysis of technology,** and this reveals philosophical prejudices and assumptions in technology, reconstructs accepted philosophical concepts in technology and engineering, and clarifies any unclear use of concepts. (C) Philosophy in technology analyzes **the consequences of philosophical prejudices in technology,** thus revealing the philosophical assumptions in technology, determining their role in specific technical realizations, and analyzing the consequences and possible postulates for changes in philosophical underpinnings.

5. Exemplary study cases

In the light of philosophy in technology, problems that seem exclusively technical instead emerge as multidimensional concepts that draw on ideas from ontology, metaphysics, philosophy of mind, and philosophy of nature. To exemplify this, the following cases reveal how sticking exclusively to the technological perspective acts as a limiting factor on technology itself by enforcing a myopic vision of technical endeavors, thus constraining the range of possible solutions, much to the detriment of technology itself.

Searle's Chinese Room: This argument concerns whether modern computers, as based on a Turing machine (TM), operate on a syntactic or semantic level. If Searle's argument holds, computers will never think like humans, so it is impossible to base artificial general intelligence (AGI) on a TM. The insight from philosophy in technology here, namely "what is wrong with this discussion," does not come from technology but rather from philosophy of mind. It asserts that the discussion is misguided in that a TM model of the human mind would be inappropriate because if an AI system were to match human-level intelligence (i.e., AGI), it would need to be embodied, embedded, extended, and enactive (i.e., embodied cognition). Indeed, such a system appears to be our best bet so far for modeling the mind. Thus, Searle's Chinese room argument is correct in the sense that computers will never think like humans because the mind is not a TM. Nevertheless, at the same time, Searle's argument is wrong in that

the issue is not whether TM systems understand semantics in addition to syntax, which can be used to argue against large language models (LLMs) like ChatGPT 3/4 and such like. Instead, the real issue is whether TM-based models would simply be incorrect models of the mind. Discussions about whether Searle was right or not are misdirected, yet they still retain their vigor. (For an extended discussion, see, for example, the works of Smith [1998, 2019], Dreyfus [2016], Cole [2020], and Woodridge [2020].)

Synthetic phenomenology: Some in the philosophically minded AI community, such as Thomas Metzinger, have proposed a global moratorium on synthetic phenomenology until 2050, although this deadline could be revised. Synthetic phenomenology aims to model, design, and develop conscious systems, including their states and functions, using artificial hardware. The philosophical objections to the development of synthetic phenomenology are twofold: (1) Entities, whether artificial or natural, that have consciousness or phenomenal experience will also have the capacity to suffer, and we should avoid creating additional entities that would increase the overall level of suffering among conscious beings. (2) Creating entities with an artificial consciousness will essentially create “alien” beings, and we cannot predict how they will act, what moral code they will adopt, or how they will perceive us humans. It seems that engineers do not spend much of their time on these aspects of synthetic phenomenology, but we think they should. Stuart Russell’s (2019) question about AI (“What if AI succeeds?”) therefore remains open. (For an extended discussion of this, see, for example, the work of Arabales et al. [2009], Chrisley [2009], Metzinger [2021], Alexander [2022], and Cali [2022].)

Orthogonality thesis and AGI: AI systems still cannot replicate human intelligence and a human agent’s ability to cope with reality. One of the cited obstacles here is the orthogonality thesis, which holds that “the final goals and intelligence levels of artificial agents are independent of each other.” This is obviously not a computing problem, because the orthogonality thesis is a philosophical concept from philosophy of mind. After limiting the discussion of the orthogonality thesis to the ethical dimension, it could be restated as “the level of intelligence of an AI agent does not correlate with its ethical capacities.” Our experiences with psychopaths provides strong confirmation of this. “So what?” you may ask. The philosophical insight here is that the orthogonality thesis indicates that human intelligence is a complex of several relatively independent but connected faculties, and developing rational synthetic intelligence (e.g., AGI) will not automatically create moral systems. (For an extended discussion of this, see, for example, the works of Brooks [1991], Minsky [1991], Armstrong [2012], Dreyfus [2016], Wooldridge [2021], Smith [2018], Boltuc [2020], Mitchell [2020], and Roitblat [2020].)

AI and Ethics: Today’s ethics in AI systems (e.g., robots, bots, etc.) are implemented using the abstract computational model of the Turing machine (TM), much like all computer software currently is. From the ethical perspective, AI systems based on the TM model can only be behavioral systems. Engineers say that this will do for them, at least for now, as they continue to elaborate solutions to harebrained ethical problems like the trolley problem. Nevertheless, the behavioral approach to morality has been long discounted, perhaps with the exception of Big Five. Thus, TM-based bots and such like implement the wrong ethical model under the wrong computational paradigm. The philosophical insight here is that one of the critical differences, although not the only one, between synthetic ethical systems, as they are currently designed, and those of humans lies in their ethical decision-making methods. Higher levels of ethical proficiency in AI systems could be realized by adapting the Aristotelian concept of phronesis, which introduces the concepts of the *telos* of ethics (i.e., the ultimate goal of ethics) and*eudaimonia* (i.e., the best good). In the case of AI systems, the telos of their ethical decisions should be the best good of the human actors involved rather than the AI entities themselves. (For a lengthier discussion, see the works of Aristotle [2004], Wallach [2004], Wallach et al. [2010], Leslie [2019], Russell [2019], Coeckelbergh [2020], Polak and Krzanowski [2020], Dubber et al. [2020], Powers and Gansacia [2020], Muller [2021], and Veliz [2021].)

Whole brain emulation: A whole brain emulation (WBE) recreates a fully functional brain, such that it is functionally indistinguishable from the original mind. We are of course not at this level of technology yet, and no one knows if or when we will ever be there. Nevertheless, is WBE at least a tenable idea? How can we possibly answer this question? The route to an answer may come from philosophy. As it is currently conceptualized, WBE is based on eight assumptions. For example, the third assumption is that the relevant functions of the brain are Turing-computable, while the seventh assumption states that at the emulated level, the simulated components can be realized in an operational computer (i.e., a TM). The eighth assumption is that while a WBE must reproduce the original brain’s functions, it need not necessarily replicate all of the body’s other functions, so a WBE does not need to be a fully embodied brain. Now, these assumptions are philosophical rather than technical. They assume that there is a specific model for the mind/brain, one based on physical reductionism, such that a TM can host a computational model of the mind. The probity of such models needs to be addressed before we can pass judgment on the whole WBE project. Thus, the answers to the question of WBE’s feasibility lie in philosophy of mind, phenomenology, neurology, and so on rather than in technology. Philosophy does not have these answers yet,

but neither does technology. (For an extended discussion, see, for example, the works of Koene [2006, 2012, 2013], Sandberg and Bostrom [2008], Sandberg [2013], and Shanahan [2015].)

Meta-ontology in robotics: The AI systems we currently design and implement cannot replicate a human agent's ability to cope with reality. One of the reasons for this failing is, it seems, that these AI systems lack a proper ontology or representation of the real world. The computational sciences redefine what philosophical ontology, or the representation of the world, is about, because it loses its essence to the question of "what is" by focusing instead on abstract internal representations. Thus, what sort of ontology should AI systems have in order to replicate the ability of human agents to cope with real-life situations? This is a meta-ontological question. Such AI systems should have an ontological commitment to the real world, and the truth condition (ontology) of these AI systems should be consistent not with internal AI-based ontological theories but rather with the real world, such that it can be verified through operational success. (Operational or performative success refers to the ability to cope with specific tasks.) Finally, the ontology of these AI systems must account for the dynamic environment of the real world rather than using static internal representations of the world. (For an extended discussion, see, for example, the works of Brooks [1991], Berto and Plebany [2015], Dreyfus [1991], Hutchins [1995], Minsky [1991], Roitblat [2020], Smith [1998, 2019], and Krzanowski and Polak [2022].)

6. Conclusions

The list of philosophical problems in technology presented here is selective, limited, and biased toward computing technology and AI, which are the hot topics of today. Nevertheless, this limited selection should not be interpreted as implying that philosophical problems in technology are specific to just these technologies.

The lessons we can draw from this discussion are as follows:

1. Technology tends to substitute its own meaning for terms with traditional connotations in philosophy. For example, synthetic ethics are not ethics as we generally understand them, and synthetic phenomenology is not phenomenology. Likewise, synthetic ontology is not ontology, and an autonomous agent is not an agent, nor is AI actual intelligence, and so on. As such, there is a wide gap between what engineers claim to have done and what they have actually achieved. The differences in these technological and philosophical concepts are often so significant that they may refer to completely different things, such as with ethics, ethical behavior, justice, agency, autonomy, intelligence, the mind, and so on. In this way, the meaning of such terms becomes obfuscated and lost, having taken on a rather limited and narrow interpretation. PinT therefore needs to reopen the lost semantic horizon.
2. Semantics matter: The meanings we attribute to specific terms like ethics, justice, the mind, intelligence, phenomenology, and so on matter, because they define the horizon of research and study objectives. Indeed, incorrect meanings lead to a myopic view of technology.
3. Technologists must understand that cannibalizing the deep philosophical meanings of many concepts will only work to their own detriment, while philosophers need to realize that their reflections are simply perceived as musings that lack any relevance to technology. In reality, both sides are wrong, so both sides should seek to understand each other.
4. There should be an open dialogue where both sides (i.e., technologists with a philosophical bent and philosophers with a technological understanding) can freely exchange their ideas without fear of being shouted down as ignoramuses or simpletons.

Acknowledgements

I would like to thanks Prof. Pawel Polak for the useful discussions and ideas that contributed to this paper.

Footnotes

[1] This paper is based on a previous paper that was presented at the Workshop on Philosophy in Technology: The Philosophical Challenges for Technology from Various Points of View on April 28–29, 2023 at Wrocław University of Science and Technology. Contact : rmkrzan[at]gmail.com

References

- Aleksander, I. 2022. "From Turing to Conscious Machines" Philosophies 7, no. 3: 57https://doi.org/10.3390/philosophies7030057
- 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
- Aristotle, 2004. The Nicomachean Ethics, J. A. K. Thomson (tłum.), London: Penguin Classics.
- Armstrong, S., 2012. General Purpose Intelligence: Arguing the Orthogonality Thesis. Available athttps://www.fhi.ox.ac.uk/wp-content/uploads/Orthogonality_Analysis_and_Metaethics-1.pdf
- Berto, F. and M. Plebani. 2015. Ontology and Meta-ontology. London: Bloomsbury.
- Brooks, R.A. 1991. Intelligence without representation. Artificial Intelligence 47 (1991), 139–159.
- 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
- Chrisley, R. 2009. Synthetic Phenomenology. International Journal of Machine Consciousness 2009 01:01, 53-70. DOI: 10.1142/S1793843009000074.
- Coeckelbergh, M. 2020. AI Ethics. Cambridge: The MIT Press.
- Cole, D. 2020. “The Chinese Room Argument”, The Stanford Encyclopedia of Philosophy (Winter 2020 Edition), Edward N. Zalta (ed.), URL = https://plato.stanford.edu/archives/win2020/entries/chinese-room/.
- Dreyfus, H. L. 2016. Skillful Coping. Oxford: Oxford University Press.
- Dubber, M. D., F. Pasquale, and D. Sunit. (eds). 2020. The Oxford Handbook of Ethics of AI. Oxford: Oxford University Press.
- Hutchins, E. 1995. Cognition in the Wild. Cambridge: The MIT Press.
- Koene, R. A. 2006. Scope and resolution in neural prosthetics and special concerns for the emulation of a whole brain, The Journal of Geoethical Nanotechnology 1, 21–29. https://www.terasemjournals.org/GNJournal/GN0104/koene_01a.html
- 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., & Polak, P. 2022. The meta-ontology of AI systems with human-level intelligence. Philosophical Problems in Science (Zagadnienia Filozoficzne W Nauce), (73), 197–230. Retrieved from https://zfn.edu.pl/index.php/zfn/article/view/610
- Leslie, D. 2019. Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector. The Alan Turing Institute. https://doi.org/10.5281/zenodo.3240529
- Metzinger, T. 2021. 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.
- Minsky, M. 1991. Logical versus analogical or symbolic versus connection or neat versus scruffy. AI Magazine, 12(2), 34–51. https://web.media.mit.edu/~minsky/papers/SymbolicVs.Connectionist.html
- Müller, V. C., 2021. "Ethics of Artificial Intelligence and Robotics", The Stanford Encyclopedia of Philosophy (Summer 2021 Edition), Edward N. Zalta (ed.), URL = https://plato.stanford.edu/archives/sum2021/entries/ethics-ai/ .
- Polak, P. & Krzanowski, R. 2020. Ethics in autonomous robots as philosophy in silico: The study case of phronetic machine ethics. Logos i Ethos. 52. 33. 10.15633/lie.3576.
- Powers, T.M., and J-G., Gansacia. 2020. The Ethics of the Ethics of AI. In Dubber, M. D., F. Pasquale, and D. Sunit. (eds). 2020. The Oxford Handbook of Ethics of AI. Oxford: OUP.pp. 27-53.
- Roitblat, H. 2020. Algorithms Are Not Enough. Cambridge: The MIT Press
- Russell, S., 2019. Human Compatible. AI and problems of control. London: Penguin.

- 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.
- Shanahan, M. 2015.The Technological Singularity. The MIT Press, Cambridge.
- Smith, B. C. 1998. On the Origin of Objects. Cambridge: The MIT Press.
- Smith, B. C. 2019. The Promise of Artificial Intelligence. Cambridge: The MIT Press.
- Véliz, C. 2021. Moral zombies: why algorithms are not moral agents. AI & Soc 36, 487–497 (2021).
- Wallach W., Allen C., 2009. Moral Machines: Teaching Robots Right from Wrong, Oxford University Press.
- Wallach W., Franklin S., and C. Allen., 2010. A conceptual and computational model of moral decision making in human and artificial agents, „Topics in Cognitive Science", vol. 2, no. 3, 2010, pp. 454–485.
- Woodridge, A. 2020. The Road to Conscious Machines. London: Penguin.