Philosophy & Technology (2024) 37:56
https://doi.org/10.1007/s13347-024-00747-7

COMMENTARY

That's Why is Worth Continuing to Think About Our Successors – A Reply to Erler

Andrea Lavazza[1,2] · **Murilo Vilaça**[3]

Received: 29 March 2024 / Accepted: 18 April 2024 / Published online: 25 April 2024

The prospect of Earth becoming uninhabitable for carbon-based life forms, including us humans, presents an undesirable scenario (Lavazza & Vilaça, 2024b). However, in contemplating the inevitability of such a hypothetical circumstance, it is imperative to explore all viable strategies aimed at preserving remnants of human values in our absence. Erler (2024) asks us if it would be worth creating the AI robotic successors we proposed (Lavazza & Vilaça, 2024a). Even in the face of the criticisms raised, our response is yes, but we do not advocate that they need to be created to accelerate our extinction as we are imperfect entities. In this vein, we distance ourselves from the two extremes of the debate: those who prophesy the AI apocalypse and those who defend a soteriological perspective of AI. Put another way, we are neither detractors nor apologists for AI.

AI is what Agar (2024)[1] calls a *charismatic extinction threat*. Taking 'charismatic' to mean that AI draws attention, there is an ambiguity here. While it arouses many fears (Apocalyptic AI), too many positive expectations may be placed on it (Salvific AI). As far as we are concerned, such prospects are just as likely to be doomed to failure. We therefore propose a 'third way'.

Erler's (2024) claim that it would be more helpful to try to avoid human extinction than to produce digital successors seems compatible with our approach. As we argued, human life is full of value. We do not find it ethically attractive to make efforts to simply

[1] Agar, N. (2024). *How to set the right level of collective worry about Artificial Superintelligence*. Pre-presentation paper.

✉ Andrea Lavazza
lavazza67@gmail.com

✉ Murilo Vilaça
murilo.vilaca@fiocruz.br

[1] Department of Brain and Behavioral Science, University of Pavia, Pavia, Italy

[2] Centro Universitario Internazionale, Arezzo, Italy

[3] Department of Human Rights, Oswaldo Cruz Foundation (Fiocruz), National School of Public Health (ENSP), Rio de Janeiro, Brazil

give birth to posthumans, especially if this puts human life at risk (Rueda, 2022). Unlike what Erler (2024) states, we do not quickly accept the prospect of human extinction. We agree with Erler that it would be interesting 'to use advanced AI technology to help develop a mitigation strategy for such a tragic scenario'. However, if these efforts are unsuccessful, and human extinction proves unavoidable, our proposal would still be an alternative worth considering. The significance of our contribution is evident herein.

But, in general, we could resort to AI for both purposes at the same time. We do not think that trying to avoid human extinction and, on the other hand, developing robotic digital successors – like an insurance policy, in case the first endeavor does not work as expected – are incompatible actions. So, the most preferable option is to join our efforts in exploring all the possibilities of AI that we think are useful for preserving human value, both before and after extinction.

We have one more point of agreement with Erler. We share the concern for AIs that are safe and trustworthy. This is a central point of the current debate (cf. Floridi, 2023b). We cannot here delve into the controversy over which of the various conceptions of ethics or justice will be implemented, how to translate them in algorithmic terms, and what theoretical and empirical trade-offs (ethics/fairness versus performance/accuracy, the well-known *Pareto curves* or *Pareto frontier*, that also raise normative questions) are involved in creating 'good biases' (Kearns & Roth, 2020). We just want to point out that the issue of safety would apply both to our idea of successors and to the AIs we would develop to save us from extinction.

A rogue 'salvific AI' could identify the pattern of a threat and, instead of avoiding it, increase its destructive power, shortening our time on Earth. The example of AI that spreads cancer given by Erler (2024) seems to point to this. So, Erler's 'salvific AI' could be as unconvincing and even irresponsible as the successors we proposed. However, it's likely that none of us thinks of acting irresponsibly, even if the consequences of our proposals could be useless or disastrous.

But what is a *rogue AI*? It often sounds as if it is a conscious, autonomous entity that decides to act dishonestly. 'Put to its [computational resources] own use instead', as Erler (2024) states in fn. 1, and being rogue are very different things. Perhaps the question is better framed in terms of AI no longer being subservient to human interests rather than in terms of dishonesty. Continuing to use the term 'rogue' – which we do not usually use to refer to our pets, for example – to define an AI that does not work or act as we would like ends up committing us to something that is the subject of much controversy, namely that our successors would display some qualities similar to ours, as we discussed in our paper (Lavazza & Vilaça, 2024a).

The best salvific AI would be a perfect zombie, i.e., something that acts as designed and carries out orders 'blindly', without consciousness, freedom, or any possibility of non-instrumental intelligence. Thus, the zombie AI would not be a problem but a good solution, except that it would not preserve the value that human beings have (cf. Floridi, 2023a). But that's not our point. We think that a kind of robotic AI successors could have a value oriented towards the preservation of human heritage and the creation of new forms of life that can achieve consciousness and thus value in the full sense.

We might also consider making our successors a sort of hybrid beings. Instead of being based on silicon alone, they could develop synthetic biological intelligence or a hybrid mind resorting to human brain organoids, i.e., lab-grown 3D mini-models of the brain

with structural and functional characteristics similar to those of an adult nervous system connected with digital devices to have both sequential and parallel forms of information processing. Advancements in biocomputing research are promising in terms of shedding light on some of the key questions regarding AI: dynamic response to external/environmental stimuli (actively modifying the environment or suitably adapting to environmental changes), computational energy consumption, AI learning limitations, phenomenal consciousness, digital sentience, moral status… (Kagan et al., 2022; Smirnova et al., 2023). There are relevant differences between BrainDish and organoid intelligence, and we are aware of the ethical issues raised by the use of human brain organoids (Lavazza, 2021). Yet, by researching our potential successors, we could get closer to entities endowed with some value and useful in the face of human extinction that remains the worst-case scenario.

Funding The authors do not have specific funding to be declared.

Data Availability N.A.

Declarations

Ethics Approval and Consent to Participate N.A.

Consent for Publication N.A.

Competing Interests The authors declare that they do not have comepting interests.

References

Erler, A. (2024). AI successors worth creating? Commentary on Lavazza & Vilaça. *Philosophy & Technology*, *37*, 40.

Floridi, L. (2023a). AI as *agency without intelligence*: On ChatGPT, large Language models, and other generative models. *Philosophy & Technology*, *36*, 15.

Floridi, L. (2023b). *The ethics of artificial intelligence: Principles, challenges, and opportunities*. Oxford University Press.

Kagan, B. J., et al. (2022). *In vitro* neurons learn and exhibit sentience when embodied in a simulated game-world. *Neuro*, *110*(23), 3952–3969.

Kearns, M., & Roth, A. (2020). *The ethical algorithm. The science of socially aware algorithm design*. Oxford University Press.

Lavazza, A. (2021). 'Consciousnessoids': clues and insights from human cerebral organoids for the study of consciousness. *Neuroscience of Consciousness*, 2021(2), niab029.

Lavazza, A., & Vilaça, M. (2024a). Human extinction and AI: What we can learn from the ultimate threat. *Philosophy & Technology*, *37*(1), 16.

Lavazza, A., & Vilaça, M. (2024b). Ways of addressing human extinction – a reply to Glannon. *Philosophy & Technology*, *37*(1), 41.

Rueda, J. (2022). Genetic enhancement, human extinction, and the best interests of poshumanity. *Bioethics*, 1–10.

Smirnova, L., et al. (2023). Organoid intelligence (OI): The new frontier in biocomputing and intelligence-in-a-dish. *Frontiers in Science*, *1*, 1017235.