Philosophy & Technology (2024) 37:41
https://doi.org/10.1007/s13347-024-00733-z

COMMENT

Ways of Addressing Human Extinction – a Reply to Glannon

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

Received: 2 March 2024 / Accepted: 5 March 2024 / Published online: 11 March 2024

In our paper 'Human extinction and AI' (Lavazza & Vilaça, 2024), we propose to consider a scenario in which the uninhabitability of planet Earth for carbon-based living forms such as us humans becomes complete. This leads us to hypothesise the creation of a kind of our successors based on silicon and non-human in physical constitution, but inheriting human-like cognition and orientation. It is not a question of uploading mental content, a task that is difficult to achieve, but of instructing an artificial intelligence system so that it has specific behavioural dispositions (in addition to the most extensive knowledge base possible) typical of human beings.

Walter Glannon (2024), in his insightful commentary, first asks whether 'value could emerge from the creation of successor beings', which can rightly be considered posthumans. If intrinsic value, as many may plausibly think, is linked to sentience, non-sentient successors may nevertheless have instrumental value to an intrinsic value. They do so as active witnesses to the acts of value performed by human beings and as proactive entities in the construction of intrinsic value, whether in the form of possible technological evolutions capable of raising consciousness in machines or researching, replicating, and accelerating natural processes similar to those that naturally gave rise to conscious life on Earth.

Would such successors with silicon brains be categorically distinct from us? Recently, Mei and colleagues (Mei et al., 2024) conducted a Turing test on AI chatbots, evaluating their performance across a series of established behavioural paradigms aimed at eliciting traits such as trust, fairness, risk aversion, and cooperation. Additionally, they assessed their responses to a conventional Big-5 psychological inventory, measuring personality attributes. ChatGPT-4 demonstrated behavioural and personality characteristics that closely resemble those

✉ 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] Oswaldo Cruz Foundation (Fiocruz), National School of Public Health (ENSP), Rio de Janeiro, Brazil

of randomly sampled human participants drawn from a diverse cohort spanning numerous nations. Notably, the author claimed, chatbots also exhibited adaptive behavioural patterns influenced by prior interactions and situational contexts, akin to a learning process.

It is therefore uncertain whether artificial successors involve a transition from one living species to another categorically different one. In one sense, there is certainly a strong divide that cannot be compared to the evolution from one species to another in our natural history. On the other hand, it should not be overlooked that our successors are 'machines' programmed by humans to behave according to our better angels, so to speak (Pinker, 2011). Herein lies one of the possible advances that the idea of digital successors can get even as a thought experiment. For if we wish to build individuals with the behavioural dispositions best suited to preserving the instrumental and potentially intrinsic value of human beings, we must make a selection that is extremely problematic, subject as it is to moral disagreement. Here then, the use of Machine Learning as an emerging feature of artificial intelligence may be a way of making such a process of selecting the characteristics to be reproduced in successors more effective and less controversial.

This way is also a potential suggestion for choices we humans should make before extinction—whether near or far, we are agnostic about such a prediction. Faced with this opportunity to resort to efficient algorithms, capable of making right rational and moral decisions, the problem of 'freedom to fall' may arise, as suggested by Glannon, who wonders whether this would be a way of being superior to fallible humans. We believe that there is no such thing as infallibility even from advanced machines, because what we are talking about with information-generated posthumans is a group of individuals with characteristics that we believe, as humans, to be among the best we currently possess. Such digital successors would then begin to interact with each other in a modified and unpredictable environment—remember that it would be an environment uninhabitable by humans.

Since it would be desirable to allow successors degrees of behavioural freedom so that they could cope with the changing environmental conditions, the interaction among posthumans and between posthumans and the environment is as unpredictable as the outcomes in terms of individual behaviour. Our successors would thus be potentially superior to us at least under the initial conditions—they would be devoid of some of those selfish and aggressive tendencies that we carry as a biological inheritance of our evolution, and they would be immune to some cognitive biases—but this does not guarantee that they would develop an ideal society, whichever way you look at it.

The point of our thought experiment is that even today, in the face of epochal crises, we do not find a consensus on what are the best personal dispositions to which we should resort to deal with the most serious emergencies. It does not seem, for example, that on average we are sufficiently aware and proactive in the face of climate change that threatens our existence as a species. Some have controversially proposed some forms of compulsory moral enhancement (Persson & Savulescu, 2012; cf. Vilaça & Lavazza, 2022; Lavazza, 2019). We think that the use of artificial intelligence support in decision-making to overcome disagreement and achieve more effective outcomes is a possibility we need to explore.

We agree with Glannon, finally, on one point: relying on machine learning cannot give us any guarantees that the result would lead to 'perfect' or at least much better successors than us, since they would be based on human concepts in all relevant respects. And on the other hand, if they were totally different entities, human normative concepts would not apply to them. And it is true that there is no objective 'view from nowhere' providing a universal model of normativity. However, throughout human history we have pursued moral progress that has led to some undeniable advances (Buchanan & Powell, 2018). The purpose of relying on more powerful and performant tools lies in the attempt to accelerate that moral progress, especially in the face of particularly pressing challenges. We will never have perfect entities, but this should not discourage us from continuing to look for new ways to improve ourselves as individuals, our species, and successors (even non-carbon-based) that share certain values and goals. In this sense, the value that human beings embody can be maintained, albeit in different and perhaps currently unimaginable but not necessarily incommensurable forms.

Funding The authors do not have any funding to declare.

Declarations

Ethical Approval Not applicable.

Consent to Participate Not applicable.

Competing Interests The authors declare that they do not have any conflicts of interest.

References

Buchanan, A., & Powell, R. (2018). *The evolution of moral progress: A biocultural theory*. Oxford University Press.

Glannon, W. (2024). Commentary on "Human extinction and AI: What we can learn from the ultimate threat." *Philosophy & Technology, 37*(1), 26.

Lavazza, A. (2019). Moral bioenhancement through memory-editing: A risk for identity and authenticity? *Topoi, 38*(1), 15–27.

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

Mei, Q., Xie, Y., Yuan, W., & Jackson, M. O. (2024). A turing test of whether AI chatbots are behaviorally similar to humans. *Proceedings of the National Academy of Sciences, 121*(9), e2313925121.

Persson, I., & Savulescu, J. (2012). *Unfit for the future: The need for moral enhancement*. Oxford University Press.

Pinker, S. (2011). *The better angels of our nature: Why violence has declined*. Viking Books.

Vilaça, M., & Lavazza, A. (2022). Not too risky. how to take a reasonable stance on human enhancement. *Filosofia Unisinos*, *23*(3), https://doi.org/10.4013/fsu.2022.233.05