This version of the article has been accepted for publication and is subject to Springer Nature's Accepted Manuscript terms of use, but is not the Version of Record and does not reflect post-acceptance improvements, or any corrections. The Version of Record is available online at: https://doi.org/10.1007/s13347-024-00732-0

AI Successors Worth Creating? Commentary on Lavazza & Vilaça

In their article "Human Extinction and AI: What We Can Learn from the Ultimate Threat" (Lavazza & Vilaça, 2024), Andrea Lavazza and Murilo Vilaça put a new twist on the provocative idea, previously suggested by authors such as Hans Moravec (Moravec, 1988) and Marvin Minsky (Minsky, 1994), that the replacement of human beings by artificial, AI-based successors could represent a desirable historical development, at least if certain conditions are met. Lavazza and Vilaça's proposal distinguishes itself in the way it spells out those conditions: the authors do not describe such replacement as an ideal scenario, but rather as a form of insurance policy against certain forms of existential risk – more precisely, as a fallback option in the event of a catastrophic outcome for humanity, triggered by phenomena like runaway climate change or a highly lethal and out-of-control pandemic.

The authors' rich analysis raises a number of interesting questions regarding what counts as a responsible approach to existential risk, the nature of personal identity, and what exactly gives human life its value. One concern that comes to mind about their proposal, however, is that it involves being too quick to accept the prospect of human extinction, choosing to use advanced AI technology to help develop a mitigation strategy for such a tragic scenario (tragic in the authors' own view as well), rather than focusing single-mindedly on avoiding it, including with the help of AI. It might be argued that the efforts and resources that would be required by the kinds of preparation suggested by the authors would be better spent towards avoid human extinction in the first place. These efforts, one might add, should also involve work on AI safety, since AI itself, as the authors mention, represents a possible future source of existential risk alongside the other ones they cite, a source that the authors' own proposal may not be able to deal with, given that a rogue superintelligent AI could plausibly interfere with the plan to create artificial successors.

The authors do respond that whether or not such interference would occur would depend on whether a network of AI successors were ready to be deployed at the time the superintelligence went rogue, and were able to counteract its destructive impact. Such a possibility, however, would seem to entail that the issue of AI safety had been successfully resolved: a "friendly" AI system, or network of such systems, would have been developed with the capacity to stop the emergence of an out-of-control and destructive superintelligence. If this friendly AI were able to prevent any interference from a rogue superintelligence in the creation of artificial successors, we would also expect it to be able to forestall human extinction from the same cause. The authors are correct that even a rogue AI that caused the destruction of all humanity need not necessarily seek to also prevent the emergence of our AI successors. As they point out, such scenarios remain highly speculative, and it is difficult to anticipate what specific goals such a system would pursue.[1] Still, if the decision whether to

[1] Although, according to Nick Bostrom, any superintelligent AI would be highly likely to include the acquisition of ever greater computational resources among its instrumental goals (Bostrom, 2014). If that is correct, then one might expect that a rogue superintelligence would be motivated to interfere with the implementation of the artificial successors envisaged by Lavazza and Vilaça, since they would require computational resources that it could put to its own use instead.

interfere or not were left to the rogue AI itself, this would not represent a satisfactory insurance policy against the tragedy of human extinction.

The authors also suggest that a humanity-ending superintelligent AI might itself, depending on its exact characteristics, be viewed as a reasonably satisfactory, if not ideal, successor to humans. While this is not inconceivable, it nevertheless seems unlikely under the standard risk scenarios involving machine superintelligence, in which a typically non-conscious but very powerful AI relentlessly pursues a goal we ourselves may initially have specified, but in perverse ways we had not anticipated – say, infecting the whole of humanity with cancer to help find a cure as fast as possible (Russell, 2019). There would seem to be little reason to expect that such a system, which had been designed solely as a highly effective tool for human use (rather than to have intrinsic value or moral status), would display the qualities that Lavazza and Vilaça expect from our AI successors.

That being said, one might reasonably reply on the authors' behalf that the idea of solely focusing on the avoidance of existential risks, without considering any mitigation strategies that might be employed in a worst-case scenario, is not compelling, and is even irresponsible. While we may of course hope that we will be able to indefinitely keep such risks at bay, we cannot be certain that we will succeed. Given that uncertainty, one might think it wise to ponder the sort of "insurance policy" outlined by Lavazza and Vilaça. While this is a fair response, we still need to ask: assuming their proposal were successfully implemented, how valuable would the resulting post-extinction world be?

Seeking to answer that question requires addressing fascinating yet difficult issues about the possibility of machine consciousness, personal identity, and the nature of human well-being, issues that have already preoccupied philosophers for decades, and even centuries. In section 3 of their paper ("How to Create Our Successors"), the authors seem to envisage the possibility that our AI successors might be psychologically continuous with us, even if they lack phenomenal consciousness (i.e. subjective experience). While this may sound like a strange idea, it does not seem incoherent. Indeed, psychological continuity, as traditionally defined in the contemporary philosophical literature, does not imply an uninterrupted stream of consciousness (which humans do not possess anyway), but rather a sufficiently robust chain of connections between an individual's mental states (including their beliefs, desires, perceptions and memories) at different points in time (Parfit, 1984). Assuming it is possible to define the relevant mental states in purely functional terms, so that they need not necessarily be accompanied by any phenomenal qualities or "what-it-is-like" character (admittedly a controversial assumption),[2] we are led to contemplate the surprising possibility that we could in fact survive as the "AI zombies", or unconscious AI systems that Lavazza and Vilaça envisage as our successors.

Even more radically, if such assumptions are correct, it could in fact be *good* for us to continue existing in this form. For this to be the case, a further presupposition would need to be introduced: namely, the claim that some human goods are not fundamentally tied to subjective experience, contrary to what hedonists about well-being believe. Such distinct goods might include the satisfaction of our desires, the attainment of virtue, or of certain

[2] Some might suggest that even perceptual experiences could in principle be understood solely in terms of information processing and associated behavioural dispositions, with no necessary connection to phenomenal consciousness or "qualia", so that even the philosophical "zombies" imagined by David Chalmers could have such experiences (Chalmers, 1996). Of course, the very possibility of such zombies is itself a contested issue in the philosophy of mind.

kinds of knowledge and excellence, and valuable relationships with others (for an overview of the issue, see Crisp, 2021). On this view, such goods would also be within the reach of our AI successors, and therefore, assuming we were psychologically continuous with them (and that the psychological continuity view of identity is correct), we ourselves could continue to enjoy them even in a post-extinction world. We could, paradoxically, survive human extinction!

While this is no doubt an interesting conceptual possibility, it nevertheless relies on disputed (and, I would argue, not fully persuasive) philosophical assumptions. Furthermore, even if we accept that we could survive as unconscious AI systems, and that the goods just outlined would be accessible to such systems, it is not clear that such goods would have substantial value in the complete absence of subjective experience. How many of us would choose to significantly extend our lifespans, if we were told that our extra years would be spent in a condition akin to that of a sleepwalker, utterly unaware of what we were doing, even if it meant achieving encyclopaedic knowledge or extraordinary skill at writing, chess, or piano-playing? And even if we could thereby complete projects, such as an ambitious novel or philosophical volume, that would otherwise have gotten interrupted – but neither we nor anyone else would ever get to consciously peruse our writings? I submit that few of us would want to continue our lives in such a condition.

Similarly, if we do not assume identity with our AI successors, their absence of phenomenal consciousness would cast doubt on the idea that their existence would have significant value, whether for themselves or in itself. (Consider a closely related scenario: is there really much value to be found in a world populated by powerful, yet presumably non-conscious AI systems like DeepMind's AlphaZero, and its potential equivalents in the scientific, moral, and cultural domains, all continuously taking their respective fields to greater heights and some of them general enough in their capacities to interact with one another, but with no conscious observer to take notice of these activities?) If so, it is not so clear that we would have strong reasons to try and secure such value by implementing the proposal advocated by Lavazza and Vilaça.

Our reasons to do so would be much stronger if we could assume that our AI successors would enjoy phenomenal consciousness. Therefore, Lavazza and Vilaça's line of argument could be reinterpreted as advocating for the development of conscious AI systems, ones that could produce successors with the ability to preserve at least much of what is valuable about our own existence – if not our own existence itself. If they could achieve the latter, this would entail a "benign" form of human extinction, one in which humanity disappeared, yet the individuals that originally constituted it did not (unlike in more standard, "malignant" forms of extinction), because they had transitioned to a new, "posthuman" species, as advocated by transhumanist thinkers. And even if humans could not survive in such a form, the value, both intrinsic and instrumental, of a world populated by conscious AI systems that had inherited our better qualities would be less open to question. Taking Lavazza and Vilaça's arguments seriously, then, means highlighting the timely nature of discussions around the paths towards engineering consciousness in an AI system, the kind of tests that might allow us to establish the presence of such consciousness (Schneider, 2019), and the potential ethical pitfalls of seeking to create such systems – significant implications, to be certain.

References:

- Bostrom, N. (2014). *Superintelligence : Paths, Dangers, Strategies*. Oxford University Press.
- Chalmers, D. J. (1996). *The Conscious Mind : in Search of a Fundamental Theory.* Oxford University Press.
- Crisp, R. (2021). Well-Being. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2021 edition). Retrieved March 3, 2024, from https://plato.stanford.edu/entries/well-being/.
- Lavazza, A. & Vilaça, M. 2024. Human Extinction and AI: What We Can Learn from the Ultimate Threat. *Philosophy & Technology,* 37**,** 16. https://doi.org/10.1007/s13347-024-00706-2
- Minsky, M. (1994). Will Robots Inherit the Earth. *Scientific American,* 271**,** 109-113.
- Moravec, H. P. (1988). *Mind Children : The Future of Robot and Human Intelligence*. Harvard University Press.
- Parfit, D. (1984). *Reasons and Persons*. Clarendon Press.
- Russell, S. (2019). *Human Compatible: Artificial Intelligence and the Problem of* Control. Viking.
- Schneider, S. (2019). *Artificial You : AI and the Future of your Mind*. Princeton University Press.