Qeios

Peer Review

Review of: "Bridging Today and the Future of Humanity: AI Safety in 2024 and Beyond"

David Manheim[1]

1. Technion - Israel Institute of Technology, Israel

The paper, Bridging Today and the Future of Humanity: AI Safety in 2024 and Beyond, explores AI safety within a long-term technological and societal framework. It argues that while current AI safety efforts focus on immediate concerns (such as jailbreaking, privacy, and adversarial robustness), they may not be aligned with the broader trajectory of AI development and human civilization. The author envisions a future where AI is deeply integrated into society through "the Internet of Everything," facilitated by advancements in brain-computer interfaces, intelligent chips, and breakthroughs in fundamental physics and energy generation. The paper then outlines how safety could move in the direction of supporting that future.

The approach does tend towards considering systemic issues to a greater extent than many other discussions of AI risks, which is commendable. However, it seems implausible that AI progress will occur in ways that enable the approaches proposed to be successful, but the ideas in the paper are valuable. Unfortunately, as a future forecasting and risk analysis paper, the approach is deeply flawed. In large part, this is because it commits to one envisioned future.

The paper does something conceptually related to backcasting (Holmberg and Robert 2000), where a single desired future state is considered and existing trends are being challenged. However, it fails to consider the necessary strategy for actually achieving that outcome. It does correctly critique other methods as being even more narrowly focused and reactive. On the other hand, it assumes away the most serious risks.

By focusing on the desired future, it does not consider other more appropriate tools, such as scenario analysis, where different possible futures are considered. By assuming slow adoption of AI and parallel evolution of different technologies, it treats AI alignment as an evolving process linked to societal and

 doi.org/10.32388/S8E5K9

technological shifts, rather than a potential abrupt transition. Similarly, it does not consider whether the assumptions being made that would make the proposed approach work are true, or how they might fail, and what should be done. This class of approach, Assumptions based planning, (Dewar 2002) would also be appropriate as a way to navigate towards a desired future.

To address the paper's claims more concretely, I will note a few specific questionable points. The implicit claim that current AI is not already past the "rookie level" seems mistaken. I also find the assumption that there will be no AGI before neural augmentation puzzling. This seems farfetched, given the relative maturity of the two technologies. It also seems unclear why the revolutionary era couples power generation with AI progress, and that this precedes more universal adoption of robotics and complex interactions of AI systems, a trend which already has started.

The paper also does not seriously engage with AI x-risk, AGI misalignment, economic disruptions, or geopolitical risks, and while listing and summarizing some portion of the safety literature, it does not engage with the arguments for why the contemplated longer-term risks are critical, and instead elides them. It also fails to address what I view as the underlying cause of the failures, namely, a lack of safety culture in the industry. (Manheim, 2023)

To conclude, the paper does a number of valuable things and works well as a criticism of many short-term safety viewpoints. I also think that further exploration of this as a single scenario could be valuable, even though I personally view it as implausible. Unfortunately, the paper is methodologically ill-suited to achieve its goals of providing a viable new approach to safety for AI more generally – especially because it is dismissive of so much of what makes safety thinking useful, namely, the ability to withstand the unexpected.

References

Holmberg, John, and K-H. Robèrt. "Backcasting—A framework for strategic planning." International Journal of Sustainable Development & World Ecology 7.4 (2000): 291-308.

Dewar, James A. Assumption-based planning: A tool for reducing avoidable surprises. Cambridge University Press, 2002.

Manheim, David, Building a Culture of Safety for AI: Perspectives and Challenges. SSRN, June 26, 2023. http://dx.doi.org/10.2139/ssrn.4491421

Declarations

Potential competing interests: No potential competing interests to declare.