Qeios

Peer Review

Review of: "Be Aware the Perils of Solutionism in AI Safety"

Micha Elsner[1]

1. The Ohio State University, USA

This is a short position paper arguing that current discussions on “AI safety” (loosely, the discussion about risks stemming from AI with near-human or super-human cognitive capabilities) are dominated by a search for technical solutions that ignores the larger social perspective.

Because of its length, the paper doesn't try to situate itself with respect to the literature (which I think is fair because there isn't space). However, the concept of solutionism is fairly present in the AI/AI safety debates (workshop on "Resisting AI Solutionism: Where Do We Go From Here?" Reyes-Cruz et al 2025; Lindgren and Dignum “Beyond AI solutionism: toward a multi-disciplinary approach to artificial intelligence in society" 2024, etc.), and without the author's own statement of the contributions, it can be a bit difficult to see what the paper adds beyond existing sources.

The paper offers three arguments: that AGI is not inevitable, that AI safety research should be separated institutionally from stakeholders who are invested in AI development, and that summits on AI safety should be open to a larger set of stakeholders. Of these, I think the second and third points are the most interesting, as the first has been made elsewhere and leads to a fairly generic call for regulation and resistance.

The second point is, I think, the most interesting part of the paper; I had not really considered the ways in which research on AI safety might be constrained by its institutional context. I do wonder whether the research that the author calls for is in fact already happening, but outside the context of “AI safety”---many prominent critics who think the development and deployment of these systems is a mistake also shun the term “AI” (except perhaps for the phrase “AI hype”: see, e.g., “Misrepresented Technological Solutions in Imagined Futures: The Origins and Dangers of AI Hype in the Research Community”, Thais 2024). The paper could use a few more sentences to discuss whether the issue is a lack of AI-critical research, a disciplinary boundary drawn to artificially separate two bodies of research that ought to be

 doi.org/10.32388/QS7BS0

considered together, attention paid by decision-makers only to some research and not others, or some combination of these.

The third point is also interesting, but I think it should be written to make it clear that the three recent summits under consideration are case studies of some more general phenomenon. As currently written, I think the paper runs the risk of dating itself.

Declarations

Potential competing interests: No potential competing interests to declare.