Qeios

Peer Review

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

Lotje Siffels[1]

1. Radboud University of Nijmegen, Netherlands

Thank you for the opportunity to review this commentary. I believe the article is well-written and makes a valuable contribution to the discussion on AI. The commentary highlights several key points for the ethical discussion on AGI: how the frame of its inevitability precludes a genuine ethical discussion; how evaluations of AI focus exclusively on mitigating risks; and demonstrates that it is the AI developers themselves who currently hold the most power in the ethical discussion. The main issue with the article as it stands is the overall argument. Below, I point to some things that may help to improve the article.

The article warns of the perils of solutionism and identifies three key symptoms of solutionism: an assumption of AGI's inevitability, the neglect of power dynamics, and the fallacy of technocracy. Though these three perils are described well and make a significant contribution to the ethical debate on AGI, it is not made evident how these perils are functions of *solutionism* specifically. The problem of assumed inevitability is a problem of technological determinism. It is unclear in the article what the connection is between technological determinism and technosolutionism. This link exists, and I can imagine it would be that if we have technologies that construct societal problems to solve, we do not question the development of these technologies themselves, just as we would fail to question them when we assume the technology to be inevitable. But in both cases, ethical reflection is precluded differently, and this needs some working out.

A similar problem arises in the two other sections. The excellent section on how AI safety is simply a way for major technology companies to hold power over the evaluation of their own technologies (the fox guarding the henhouse) is not clear in how this is a symptom of *solutionism.* It is also not entirely clear whether this problem is essential to the AI Safety frame, or whether a similar frame, but applied through evaluation performed by truly independent actors with no ties to the AI developers, would work.

doi.org/10.32388/QZXB44

In the third section, concerning technocracy, an important point is made that summits like the UK AI Safety Summit should include more stakeholders and fewer technocrats. Again, it is somewhat unclear how this is a symptom of solutionism. This section could also be aided by giving a bit more explanation of what these summits are, how they are set up, and what exactly their decision-making power entails.

Most importantly, I feel that the concept of (techno)solutionism is underdeveloped. It is not really defined except on p. 2 as 'a focus on solving narrowly defined problems without critically examining the assumptions that frame these problems'. It is unclear where this definition came from, and it does not align with my understanding of the concept. If you want to use the concept of technosolutionism, you might want to refer to Morozov's *To Save Everything, Click Here* (2013). And though I don't like referring to my own work in a review (and this more public forum makes this even more awkward), we put some work into distinguishing a "thin" and a "thick" conception of technosolutionism that may help as well (see: Siffels & Sharon, 2024).

I think the commentary would be aided most by deciding on one of two things: either it sticks to the main argument of being a warning about the perils of solutionism, in which case it would need a conceptualization of solutionism and how these perils follow from solutionism. The second option would be to change the main argument. The commentary may be less a showcase of the dangers of technosolutionism and more a call for the democratization of the debate concerning AGIs. All three perils seem to support this argument. Avoiding both technosolutionism and technological determinism is a necessary first step (we do, however, need more thought on how to recognize and prevent technosolutionism). Paying attention to the power dynamics in the debate and the inclusion of stakeholders similarly seems to support this argument. Another option would be to frame the article as a commentary on the 'AI Safety' frame. In that case, it would be good to first identify this frame, argue that it is the leading frame regarding ethical reflection on AI, and then argue that we need a better frame to improve it.

Otherwise, I feel that despite this being a short commentary, it offers some valuable suggestions for improving the debate on AGIs.

Declarations

Potential competing interests: No potential competing interests to declare.