Commentary

Be Aware the Perils of Solutionism in AI Safety

Yulu Pi[1]

1. University of Warwick, United Kingdom

Abstract. This brief commentary critiques dominant paradigms in AI safety research, warning against the risks of techno-solutionism in the framing and governance of artificial general intelligence (AGI). It identifies three core concerns: the presumption of AGI's inevitability, the neglect of institutional power dynamics shaping research agendas, and the over-reliance on closed expert communities in governance processes. It calls for a more inclusive, reflexive approach to AI safety that questions foundational assumptions, democratizes decision-making, and broadens the scope of legitimate research inquiry.

Advanced artificial intelligence (AI) has generated intense concerns, particularly regarding its potential to lead to general-purpose AI systems, with artificial general intelligence (AGI) as the prime form, capable of surpassing human performance across nearly all cognitive tasks. Some of the most respected AI experts warn about the existential threats posed by AGI, highlighting the profound uncertainties and risks it could bring[1][2]. In response to these challenges, the field of AI safety has emerged. Although many have cautioned that there is no definitive and widely agreed definition of AI safety as a field, the focus has largely been on advanced or frontier AI[3]. Central to this focus is the problem of alignment, which is ensuring that advanced AI systems' goals and behaviors align with human values and intentions[3][4]. However, AI safety as a field must remain vigilant against the **solutionism trap**[5], particularly techno-solutionism, the tendency to default to technical solutions for complex problems without examining underlying assumptions, power structures, and systemic factors that shape these challenges, or questioning whether the framing of these problems itself requires scrutiny. In light of these concerns, I identify three key symptoms of solutionism that AI safety should address more attentively: **The assumption of AGI's inevitability; The neglect of power dynamics. The fallacy of technocracy.**

 doi.org/10.32388/TJKH93

AI safety should not treat AGI as Inevitable

The belief that AGI is inevitable exemplifies the "solutionsism" trap in AI safety, a focus on solving narrowly defined problems without critically examining the assumptions that frames these problems. Solutionism manifests itself in a preoccupation with mitigating the hypothetical risks of AGI while treating its arrival as a foregone conclusion. This perspective obscures more fundamental questions: Is AGI a desirable goal? Does its development align with broader societal interests? If the public deem it undesirable, do we possess mechanisms to prevent its advancement? The solutionist mindset necessitates a paradigm shift toward a more critical and fundamental approach to AI safety. The historical trajectories of nuclear weapons and human embryo research showed us how societies have grappled with the legitimacy, purpose, and consequences of transformative yet ethically charged technologies. In both cases, international treaties and ethical frameworks emerged, not merely to regulate application but to place boundaries on research and development itself[6].

Similarly, AGI research demands equally rigorous scrutiny. Society must move beyond a narrow focus on risk mitigation to engage in participatory discussions that confront the deeper questions: Should AGI represent the direction of progress we collectively endorse? What aspects are we willing to accept as part of this trajectory? At the heart of this issue lies the way AGI is conceptualized. Definitions from leading AI organizations, such as OpenAI's framing of AGI as systems that "outperform humans at most economically valuable work,"[7] often embed narrow assumptions about intelligence, emphasizing economic productivity while sidelining social, cultural, and ethical dimensions[8]. Intelligence, however, cannot be reduced to economic utility alone. By emphasizing economic efficiency, these definitions skew AGI research and development toward serving the interests of those funding, controlling, and deploying the technology, often at the expense of fostering deeper human connections and collective well-being. Moreover, many definitions present intelligence as a linear hierarchy, placing AGI at the pinnacle of cognitive capability. This framing risks perpetuating harmful narratives, such as those historically used to justify inequalities based on perceived intelligence. For instance, early intelligence testing was misapplied to support discriminatory policies such as eugenics and systemic disenfranchisement[9]. A similar danger looms as AGI is valorised as "superior," potentially justifying power imbalances between humans and AI or even among human groups with or without AGI.

Crucially, the inevitability of AGI is not an empirical certainty but a narrative construction. Expert opinions diverge widely, with predictions ranging from imminent breakthroughs to outright skepticism

about AGI's feasibility[10]. This variability highlights the contingent nature of AGI's trajectory, undermining deterministic claims of its inevitability. Treating AGI as an unavoidable endpoint of progress functions as a self-fulfilling prophecy, closing off critical pathways of resistance and alternative futures.

AI safety should be sensitive to power

AI safety research, when divorced from a critical understanding of power dynamics, risks inadvertently reinforcing the very forces it seeks to regulate. The appearance of rigorous scientific oversight can coexist with a vulnerability to entrenched power structures, potentially creating a system in which the language of safety obscures deeper systemic issues.

Consider the ecosystem of AI evaluation organizations: while these entities present themselves as independent assessors, they are often deeply entangled in a web of conflicting interests and strategic positioning. The case of Dan Hendrycks, an AI safety leader simultaneously involved in drafting the California bill for the Safe and Secure Innovation for Frontier Artificial Intelligence Models Act and running a for-profit evaluation company, exemplifies how AI safety's institutional design undermines the independence of oversight organizations and governance[11]. Many prominent evaluation organizations are funded by the same sources that support seemingly independent policy institutions advocating evaluation as the primary mechanism for AI governance[12]. This creates a circular logic that ultimately reinforces institutional self-preservation[11]. Companies like OpenAI subtly boost evaluation findings that suggest advanced capabilities, using research results to implicitly support narratives of AGI's inevitability. This is a carefully calibrated strategy that generates both excitement and concern, transforming potential technological risks into a form of strategic communication and policy influences. Such tactics blur the line between genuine safety concerns and calculated efforts to influence public discourse and policymaking. As large AI corporations position themselves as guardians of safety, particularly amidst the systematic downsizing of their Trust and Safety teams[13] — one must question the authenticity of these claims. What is presented as a good-faith effort to mitigate AI risks may risk becoming a form of 'safetywashing'[14]. This aligns with a broader solutionist ethos, which, instead of interrogating the fundamental purpose of AGI, focuses on the management of narrowly defined risks, while avoiding consideration of its broader societal consequences.

A particularly concerning aspect arises when AI safety research is financially supported by the same corporations advancing AGI development[15][16]. This creates a fundamental conflict of interest, implicitly reinforcing existing power structures. The growing involvement or hiring of independent AI safety researchers by AGI-developing corporations further highlights this issue. This dynamic not only shapes the research agenda but also subtly curates which questions can be asked—and which are deemed off-limits. Funding structures and institutional frameworks do not merely finance research—they also shape its content and scope by influencing which questions are framed as legitimate or actionable within prevailing agenda. As a result, the very definition of "AI safety" becomes constrained, focusing narrowly on managing risks related to the deployment of AGI, while excluding more critical questions about the development of AGI itself. What might appear as integration into research efforts is, in fact, a mechanism of control—researchers are absorbed into the very systems they are meant to critically examine, with their independence gradually eroded by institutional logic and economic necessity.

This situation reflects the operation of power in a more insidious, invisible form: not through overt coercion, but by shaping the discourse and the boundaries of what is deemed acceptable and needed research. As such, the terms of debate are often set by powerful stakeholders whose economic and political interests are tied to the continued development and deployment of AGI, thereby limiting the space for alternative, critical perspectives. The language and practices around AI safety are themselves shaped by these power dynamics, directing attention away from the ethical implications of AGI's existence and focusing solely on its safe deployment, a focus that is ultimately rooted in the preservation of existing economic structures. A truly meaningful approach to AI safety must transcend these limitations. It requires a radical re-imagining that centers transparency, accountability, and a willingness to critically examine the underlying assumptions driving AGI.

The Fallacy of Experts Solving All Problems

UK AI Safety Summit, framed as a response to the risks posed by advanced AI systems, highlights the shortcomings of solutionism. While such gatherings convene impressive expertise, they often exclude broader public perspectives and stakeholder input. By prioritizing closed-door discussions among a select group of experts, the summit embodies a technocratic approach that excludes broader societal input, failing to engage and inform the public in a meaningful way. While events like AI Fringe 2024 in London aimed to center the perspectives of civil society, academia, and marginalized communities, the classification of these events as "fringe" rather than integral to the debate reveals a troubling hierarchy of

voices. This framing suggests that inclusive, participatory approaches are secondary to the “core” discussions controlled by technocrats. The “paradox of generative AI governance”[17] is stark: as AI becomes increasingly embedded in our lives, its governance is becoming less accessible and less democratic. The AI Safety Summit in San Francisco further exemplifies this exclusionary trend. Announced by the AI Safety Institute as an invitation-only event, the summit raises critical concerns about transparency and inclusivity. Critical questions remain unanswered: What criteria were used to select invitees? What is the process for those not invited to request participation or submit contribution?

In contrast, the AI Action Summit in France offers an alternative model, coordinating open consultations that gathered over 10,000 contributions from citizens and 200 experts worldwide. This approach highlights a strong public demand for robust, inclusive AI governance, with direct input from both civil society and technical communities. This approach demonstrated the feasibility and value of inclusive governance mechanisms while highlighting strong public demand for participation in AI policy formation. However, even this seemingly more democratic model revealed persistent limitations: the consultations predominantly reflected European and American perspectives, perpetuating geographical and institutional interests[18]. The field requires a fundamental shift from techno solutionism toward a more comprehensive approach that integrates diverse global perspectives, acknowledges power dynamics, and questions underlying assumptions about technological progress. This shift requires engaging with foundational questions I have discussed previously about AGI’s development: Is it a goal worth pursuing? How can its development align with shared societal values? Such deliberations must transcend current technocratic and corporate priorities, embracing diverse perspectives to ensure AI governance serves broader societal interests.

References

1. ^*Bengio, Y.: (2023). https://yoshuabengio.org/2023/06/24/faq-on-catastrophic-ai-risks/*
2. ^*CAIS, signatories: (2023). https://www.safe.ai/work/statement-on-ai-risk*
3. a, b*Technical report, Department for Science, Innovation and Technology and AI Safety Institute (2024). https://www.gov.uk/government/publications/international-scientific-report-on-the-safety-of-advanced-ai*
4. ^*Ji, J., Qiu, T., Chen, B., Zhang, B., Lou, H., Wang, K., Duan, Y., He, Z., Zhou, J., Zhang, Z., Zeng, F., Ng, K.Y., Dai, J., Pan, X., O’Gara, A., Lei, Y., Xu, H., Tse, B., Fu, J., McAleer, S., Yang, Y., Wang, Y., Zhu, S.-C., Guo, Y., Gao, W.: (2024). https://arxiv.org/abs/2310.19852*

5. ^Selbst AD, Boyd D, Friedler SA, Venkatasubramanian S, Vertesi J (2019). "Fairness and Abstraction in Sociotechnical Systems". Proceedings of the Conference on Fairness, Accountability, and Transparency. 59: 59–68. doi:10.1145/3287560.3287598.
6. ^Harding, V Princeton University Press (2024). https://press.princeton.edu/books/hardcover/9780691244877/ai-needs-you?srsltid=AfmBOorhRWzrvZaW3eeQg2QdS3GYWefvlH76VOK-Tb5bx3AyJN1b2cnu
7. ^OpenAI: (2024). https://openai.com/charter/
8. ^Blili-Hamelin, B., Hancox-Li, L., Smart, A.: (2024). https://arxiv.org/abs/2401.13142
9. ^McQuillan D (2022). "Resisting AI". Bristol University Press. doi:10.2307/j.ctv2rcnp21.
10. ^Science, I., Technology: Technical report, UK Government (2023). https://assets.publishing.service.gov.uk/media/65395abae6c968000daa9b25/frontier-ai-capabilities-risks-report.pdf
11. a, bLeicht, A.: (2023). https://www.antonleicht.me/writing/evals
12. ^Open Philanthropy: 2024 https://www.openphilanthropy.org/grants/?focus-area=potential-risks-advanced-ai
13. ^Firstpost: (2023). https://www.firstpost.com/tech/google-lays-off-people-trust-safety-team-responsible-ai-others-working-overtime-fix-gemini-13744836.html
14. ^Lazar S, Nelson A (2023). "AI safety on whose terms?". Science. 381 (6654): 138–138. doi:10.1126/science.adi8982.
15. ^Anthropic: (2024). https://alignment.anthropic.com/2024/anthropic-fellows-program/
16. ^Meta: (2024). https://www.llama.com/llm-evaluation-research-grant/
17. ^Ulnicane I (2024). "Governance fix? Power and politics in controversies about governing generative AI". Policy and Society. 44 (1): 70–84. doi:10.1093/polsoc/puae022.
18. ^The Future Society: (2024). https://thefuturesociety.org/consultation-interim-report

Declarations

Funding: No specific funding was received for this work.

Potential competing interests: No potential competing interests to declare.