HSPI The Open Access Publisher

JAIRI ISSN 3069-762X JOURNAL OF ARTIFICIAL INTELLIGENCE RESEARCH AND INNOVATION

Opinion

The Alignment Tax: Why Safety Shouldn't Slow Innovation

Zunaira Khalid*

Institute of Medical Engineering, Xi'an Jiaotong University, China

Abstract

The idea of the "alignment tax" often appears in discussions about developing artificial intelligence. It suggests that safety measures slow down innovation and competitiveness. This opinion piece challenges that view. It claims that safety should be seen as an important part of technological capability, not as an added cost. By looking at examples from the aviation industry and recent progress in AI research, the article shows how interpretability, constitutional AI, and scalable oversight lead to more reliable, controllable, and socially acceptable systems. It argues that the real cost comes not from investing in safety, but from ignoring it. This neglect can cause societal harm, erode public trust, and invite more regulatory scrutiny. By viewing safety as a driver of long-term innovation, this article encourages the integration of alignment research into the foundation of AI development. This approach aims for sustainable and responsible progress.

More Information

***Corresponding author:** Zunaira Khalid, Institute of Medical Engineering, Xi'an Jiaotong University, China, Email: zunairakhalid@stu.xjtu.edu.cn

Submitted: December 08, 2025
Accepted: January 07, 2026
Published: January 08, 2026

Citation: Khalid Z. The Alignment Tax: Why Safety Shouldn't Slow Innovation. J Artif Intell Res Innov. 2026; 2(1): 001-002. Available from: https://dx.doi.org/10.29328/journal.jairi.1001013

Keywords: Artificial intelligence; Alignment tax; AI safety; Responsible innovation; Interpretability; AI governance; Trustworthy AI

Opinions and perspectives

The artificial intelligence community has a false dichotomy: AI safety or AI speed. This framing implies that we have to forego innovation at the altar of conformity, or hurry on even when we are going to face existential danger. They are both problematic paths, but it continues to argue that we have no other choice than these [1,2].

The performance cost of ensuring AI systems are safe and controllable is also referred to as the so-called alignment tax, which is cited as the reason why safety research should be deprioritized. Teams cite that the introduction of safety constraints decreases model capabilities, increases the deployment timeframes, and results in a competitive disadvantage. Such reasoning is based, however, on a very limited definition of the concept of capability that values raw performance more highly than reliability, interpretability, and robustness [3,4].

Reframing the conversation

Suppose that safety research is, in fact, innovation research? Take into account the development of aviation. The invention of Flight belongs to the Wright brothers, and it was only decades of innovations in the domain of safety: pressurized cabins, redundancies, black boxes, and strict measures that led to becoming the standard of the modern aviation industry. These were not barriers to advancement, but facilitated the size and confidence to enter commercial aviation. The current AI business is at a comparable inflexion point [1,5].

More recent discoveries of mechanistic interpretability, constitutional AI, and scalable oversight are not merely safety measures but essential findings to the understanding and control of complex systems [5-7]. By figuring out how to direct the behavior of the models without re-training, or by Building systems that are able to describe the chain of their reasoning, researchers are increasing the breadth, rather than narrowing it, of AI [8-11].

The real tax

It is not the cost of construction of safe systems that is the real tax, but the cost of not building safe systems. Implemented AI systems, which simulate medical guidance, increase biases in the employment process, or create persuasive fake news, are costly to society on a colossal scale. Every failure undermines the confidence of the people, creates opportunities to regulate, and becomes a limiting factor to the future possibilities of the field, much more significantly than prudent safety work ever would [11,12].

In addition, there is a change in the competitive forces. Organizations that build strong safety practices are becoming more of leaders, rather than laggards. It is turning out to be a decisive factor, the possibility to implement AI systems that can be trusted by users, that can be accepted by regulators, and that will not lead to disastrous failures [12-15].

A path forward

The next step is to forego the safety-versus-speed frame. Rather, safety is a key capability dimension that the AI research community should adopt. It indicates investing in interpretability as we invest in scaling laws, regarding alignment research as being part of the foundation and not optional, and quantifying the success not only by benchmark scores but by aspects of reliability, controllability, and positive effects [3,13].

It does not matter whether we can afford the alignment tax. Whether we can afford not to pay it, and whether we will recognize the fact that the most valuable innovations might be those that will enable us to survive in the transformative technology we are producing [1,2,12].

This view is part of the current discussions in the field of AI safety, and it is not a consensus opinion. Other points of view focus on other trade-offs between short-term deployment imperatives and long-term safety factors.

Conclusion

The idea that AI development involves a trade-off between speed and safety is misleading and limiting. This short opinion argues that research focused on safety should be seen as a key part of innovation, not a barrier. Historical and current evidence shows that strong safety practices build trust, encourage wider use, and safeguard the long-term future of transformative technologies. By viewing safety as a crucial ability, along with performance and scalability, the AI community can shift away from chasing short-term gains and work towards lasting, socially beneficial progress. The important question is not whether we can avoid the costs of alignment, but whether we can responsibly manage the costs of ignoring safety.

Author contributions

The author solely conceptualized, drafted, and finalized the manuscript.

Acknowledgment

The author acknowledges the broader AI safety and governance research community for ongoing discussions that informed the perspectives presented in this article.

References

1. Russell S. Human compatible: artificial intelligence and the problem of control. New York (NY): Viking; 2019. Available from: https://www.goodreads.com/book/show/44767248-human-compatible
2. Bostrom N. Superintelligence: paths, dangers, strategies. Oxford (UK): Oxford University Press; 2014. Available from: https://global.oup.com/academic/product/superintelligence-9780199678112
3. Amodei D, Olah C, Steinhardt J, Christiano P, Schulman J, Mané D. Concrete problems in AI safety. arXiv [Preprint]. 2016. Available from: https://doi.org/10.48550/arXiv.1606.06565
4. Christiano P, Leike J, Brown T, Martic M, Legg S, Amodei D. Deep reinforcement learning from human preferences. Adv Neural Inf Process Syst. 2017. Available from: https://doi.org/10.48550/arXiv.1706.03741
5. Bai Y, Kadavath S, Kundu S, Askell A, Kernion J, Jones A, et al. Constitutional AI: harmlessness from AI feedback. arXiv [Preprint]. 2022. Available from: https://doi.org/10.48550/arXiv.2212.08073
6. Olah C, Cammarata N, Schubert L, Goh G, Petrov M, Carter S. Zoom in: an introduction to circuits. Distill. 2020;5(3):e00024. Available from: https://distill.pub/2020/circuits/zoom-in/
7. Leike J, Martic M, Krakovna V, Ortega PA, Everitt T, Lefrancq A, et al. Scalable agent alignment via reward modeling. arXiv [Preprint]. 2018. Available from: https://doi.org/10.48550/arXiv.1811.07871
8. Doshi-Velez F, Kim B. Towards a rigorous science of interpretable machine learning. arXiv [Preprint]. 2017. Available from: https://doi.org/10.48550/arXiv.1702.08608
9. Floridi L, Cowls J, Beltrametti M, Chatila R, Chazerand P, Dignum V, et al. AI4People—an ethical framework for a good AI society. Minds Mach. 2018;28(4):689–707. Available from: https://doi.org/10.1007/s11023-018-9482-5
10. Mitchell M, Wu S, Zaldivar A, Barnes P, Vasserman L, Hutchinson B, et al. Model cards for model reporting. Proc Conf Fairness Account Transparency. 2019:220–9. Available from: https://doi.org/10.1145/3287560.3287596
11. Jobin A, Ienca M, Vayena E. The global landscape of AI ethics guidelines. Nat Mach Intell. 2019;1(9):389–99. Available from: https://doi.org/10.1038/s42256-019-0088-2
12. Mittelstadt B. Principles alone cannot guarantee ethical AI. Nat Mach Intell. 2019;1(11):501–7. Available from: https://www.nature.com/articles/s42256-019-0114-4
13. Hendrycks D, Carlini N, Schulman J, Steinhardt J. Unsolved problems in machine learning safety. arXiv [Preprint]. 2021. Available from: https://doi.org/10.48550/arXiv.2109.13916
14. Rahwan I, Cebrian M, Obradovich N, Bongard J, Bonnefon JF, Breazeal C, et al. Machine behaviour. Nature. 2019;568(7753):477–86. Available from: https://doi.org/10.1038/s41586-019-1138-y
15. Shneiderman B. Human-centered artificial intelligence: reliable, safe, and trustworthy. Int J Hum Comput Interact. 2020;36(6):495–504. Available from: https://arxiv.org/abs/2002.04087