Helen Nissenbaum
Cornell Tech

AI Safety: A Poisoned Chalice?

We hear a lot about the awesome potential of AI—the achievements of reinforcement learning, the astonishing power of foundation models and generative AI. Amplifying the hype, AI Safety has emerged as its counterpoint. AI Safety, when I first encountered it, brought to mind autonomous vehicle crashes, nuclear meltdowns, killer drones, and robots-gone-haywire. Nowadays, I see a different, more aggressive intention as AI Safety has come to dominate the public agenda around AI, beyond the purely technical and economic.

From my perusal of publicly available material, the term appears to originate within the technical AI community, proponents of AI, including leading AI researchers, developers, entrepreneurs, and investors (for example, OpenAI's Sam Altman, UC Berkeley's Stuart Russell, and others). Shortly following, AI Safety gained traction among corporate heads, industry consortia, and dominant company policy documents. In parallel, AI Safety entered the mainstream with appropriately named academic centers, podcasts, newscasts, trade books, and popular articles. Then, in October 2023, President Biden's executive order, *Safe, Secure, and Trustworthy Artificial Intelligence*,[1] included the term safe (and variants, unsafe, safety, safeguard, etc.) approximately 60 times, compared with three times for ethics. (Severin Engelman, a DLI postdoctoral fellow, Cornell Tech, enlightened me with these counts.)

Having spent a career writing and teaching about the societal and ethical implications of computational technologies and data policies, exhorting aspiring engineers to think beyond technical properties to ethical values embedded in technical systems, you might think I would celebrate the ascendance of AI Safety. Yet, I do not. Safety and security are worthy aims, but an AI Safety monopoly[2] is a poisoned chalice.

Beguiling Rhetoric

Proponents spin the tale. AI is shockingly smart, making real a fantastical future that sci-fi creators could only conjure. On countless tasks, machines outperform any single human; in aggregate, it is not far-fetched to imagine superintelligent AGIs outperforming all humans. With a will of their own and the intellect to execute on it, embodied AGIs may no longer care to perform what we have designed them for and may even deem pesky humans unnecessary. They could pose an existential threat to humanity, or condemn us to lives of misery.

To forestall this nightmare future, proponents call for AI Safety—systems that are aligned with human purpose and human values, into which they are hardwired. This beguiling rhetoric seeks to mollify at the same time that it aggrandizes AI. Give AI creators free rein, but do not worry, they are on team human! This rhetoric anoints proponents—creators of AI systems, technologists, investors, gargantuan tech companies—as the key or only competent defenders of humanity. Like magicians, however, they draw our gaze away from a more urgent, more immanent reality.

Synecdoche: Safety Is Not Ethics

The words of Neil Postman ring in my head: "the advantages and disadvantages of new technologies are never distributed evenly among the population. This means that every new technology benefits some and harms others."[3] There are winners and losers, he would say.

The AI Safety sleight of hand is that we are "in it together," and for a narrow band of injury, we may be. For many other urgent and immanent issues, we are not. (I am not specifically referring to the appalling representational monoculture, bias, and unfair discrimination based on race, gender, etc., deservedly the focus of study and corrective policy.[4]) For many of these, AI already poses catastrophic threats unevenly across large swaths of society. For example, it presumes almost limitless data entitlements, drains the pool of common knowledge, facilitates unfair workplace and labor practices, favors the efficiency of AI operators sometimes at the

Digital Object Identifier 10.1109/MSEC.2024.3356848
Date of current version: 15 March 2024

continued on p. 94

 Copublished by the IEEE Computer and Reliability Societies

Last Word *continued from p. 96*

expense of efficiency for others, exploits people's efforts explicitly and surreptitiously, exploits environmental resources, supports manipulative practices, diminishes individual self-determination, diminishes quality of life, facilitates decisional unfairness and opacity, destroys privacy.

These effects threaten different people differently, even pitting some of us against others. Privacy, for instance, means restricting dataflows for companies; worker autonomy means imposing constraints on employers; targeted advertising means decreasing individual self-determination; AI-driven hiring systems limit opportunities for certain groups. I'm not here arguing one way or the other. The key point is that although AI Safety is a laudable mission for the narrow band of dangers threatening all humans, a preponderance of dire and immanent threats affects different people differently. The former may benefit from technocratic leadership; the latter requires measured ethical valuation.

Safety is a glove without a hand. Defined as freedom from danger, risk, or injury (see *The American Heritage Dictionary of the English Language*, 5th edition; the definition of "security" is virtually identical), safety gains legitimacy only when these terms are imbued with concrete meaning, typically, including harms such as bodily injury, death, destruction of property. Curious about the nature of the looming threats that galvanize proponents of AI Safety, I searched academic and popular writings and public websites. A common trope was the celebration of a shiny future due to astonishing advances in AI[5] and a commitment to protecting humanity against potential dangers. Beyond this trope, AI Safety proponents fell into two rough categories: 1) those who focused almost entirely on policies, frameworks, and procedures— the *who* and the *how*, but almost nothing substantive about the *what*, that is, the nature of the dangers and injuries against which AI Safety will inoculate, and the likely victims[6,7,8]; or 2) those who chronicled of long-heralded ethical concerns, including traditional safety and security issues, such as drone collisions, malicious adversaries, and buggy code, along with bias, unfair discrimination, and so on.[9,10]

The trouble is, to the extent AI Safety has dominated the public imagination, it overshadows issues that *should* dominate the public agenda, such as *which* problems AI should be tuned to address, and how to prioritize them for the fair betterment of humanity. Are there more just ways to distribute Earth's resources for these purposes, beyond the marketplace, and, also human resources? Finally, what factors, criteria, and constraints should weigh in the balance? Recognizing and resolving *these* questions is the stuff of ethics.

A Pragmatic Alternative?

If ethics can ride in on the coattails of the AI Safety hype, why not? What's in a label? This precisely inverts the relationship; ethics is not part of safety; safety is part of ethics. An ethical lens exposes diverse and often contradictory interests, purposes, and values. Ethical thinking and deliberation, guided by principles and values, finds a way through these tough questions and conflicts. Safety executes on it.

Subsuming ethics under the AI Safety label severs historical ties between contemporary concerns and age-old traditions in ethical thinking and political philosophy (inspired by Isaiah Berlin, "ethics applied to society").[11] It has already severed contemporary AI discussions from decades of accumulated research and scholarship in ethics and technology, and its subfield concerned with values embodied in digital technologies. This dooms us, at best, to reinventing the wheel. Labels matter. (A web search on AI Safety yields an OpenAI blog from 2023, a 2021 essay on Georgetown's Center for Security and Emerging Technology website, the farthest back dating to 2016). Or, worse, a not-invented-here syndrome—a form of tribalism, an unwillingness to adopt ideas that originate from another culture.[12]

Finally, an ethical lens also sees beyond individuals to societal well-being. A tragic blind spot for prior innovations in computational technologies, such as, centralized databases, the Internet and Web, social media platforms, behavioral advertising, data analytics, etc., was their corrosive impact on social institutions. In an area I know well—privacy—a sole focus on harms to individuals has been disastrous; it misses the forest for the trees. Inappropriate dataflows leach vitality from many of the institutions that sustain organized, productive societies.[13] We are still reeling from the damage of untethered algorithmic decision systems and wholesale surveillance to core institutions—education, health care, social community, employment, and, direst of all, democracy.[14] It is a mistake to reduce existential risks to humanity to aggregated risks to individuals (an abiding challenge to Utilitarian thinking), for, often, it is the integrity of societal institutions that makes human life worth living.

Key Takeaways

- When it comes to AI, we are not all in it together.
- An existential threat to social institutions is as catastrophic as threats of individual injury.
- The rhetoric of AI Safety is unmoored without a solid basis in ethics. ■

References

1. "FACT SHEET: President biden issues executive order on safe,

secure, and trustworthy artificial intelligence." The White House. Accessed: Jan. 10, 2024. [Online]. Available: https://www.whitehouse.gov/briefing-room/statements-releases/2023/10/30/fact- sheet- president- biden- issues-executive-order-on-safe-secure-and-trustworthy-artificial-intelligence/
2. S. Ahmed, K. Jazwinska, A. Ahlwat, A. Winecoff, and M. Wang, Nov. 2023, "Building the epistemic community of AI safety," SSRN. [Online]. Available: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4641526
3. N. Postman, "Five things we need to know about technological change," UC Davis, Davis, CA, USA, Talk delivered in Denver Colorado, 1998. [Online]. Available: https://www.cs.ucdavis.edu/~rogaway/classes/188/materials/postman.pdf
4. S. Lazar and A. Nelson, "AI safety on whose terms?" *Science*, vol. 381, no. 6654, p. 138, Jul. 2023, doi: 10.1126/science.adi8982.
5. "AI safety summit 2023." GOV.UK. Accessed: Jan. 10, 2024. [Online]. Available: https://www.gov.uk/government/topical-events/ai-safety-summit-2023
6. "Microsoft's AI safety policies," Microsoft, Redmond, WA, USA, 2023. [Online]. Available: https://blogs.microsoft.com/on-the-issues/2023/10/26/microsofts-ai-safety-policies/
7. "Our approach to AI safety." OpenAI. Accessed: Jan. 10, 2024. [Online]. Available: https://openai.com/blog/our-approach-to-ai-safety
8. Partnership on AI. Accessed: Jan. 10, 2024. [Online]. Available: https://partnershiponai.org/pai-model-deployment-guidance-press-release/
9. D. Hendrycks, "An overview of catastrophic AI risks," 2023. [Online]. Available:https://arxiv.org/abs/2306.12001
10. "Stanford center for AI safety," Stanford Univ., Stanford, CA, USA. Accessed: Jan. 10, 2024. [Online]. Available: https://aisafety.stanford.edu/
11. I. Berlin, *The Crooked Timber of Humanity: Chapters in the History of Ideas*. Princeton, NJ, USA: Princeton Univ. Press, 2013, p. 2.
12. "Not invented here." Wikipedia. Accessed: Jan. 10, 2024. [Online]. Available: https://en.wikipedia.org/wiki/Not_invented_here
13. H. Nissenbaum, *Privacy in Context: Technology, Policy, and the Integrity of Social Life*. Stanford, CA, USA: Stanford Univ. Press, 2010.
14. B. Laufer and H. Nissenbaum, "Algorithmic displacement of social trust," Knight First Amendment Institute at Columbia University, New York, NY, USA, 2023. [Online]. Available: https://knightcolumbia.org/content/algorithmic-displacement-of-social-trust

Digital Object Identifier 10.1109/MSEC.2024.3369438