EDITORIAL

AI safety on whose terms?

Seth Lazar
is a professor of Philosophy at the Australian National University, Canberra, Australia. seth.lazar@anu.edu.au

Alondra Nelson
is Harold F. Linder Professor in the School of Social Science at the Institute of Advanced Study, Princeton, NJ, USA. anelson@ias.edu

Rapid, widespread adoption of the latest large language models has sparked both excitement and concern about advanced artificial intelligence (AI). In response, many are looking to the field of AI safety for answers. Major AI companies are purportedly investing heavily in this young research program, even as they cut "trust and safety" teams addressing harms from current systems. Governments are taking notice too. The United Kingdom just invested £100 million in a new "Foundation Model Taskforce" and plans an AI safety summit this year. And yet, as research priorities are being set, it is already clear that the prevailing technical agenda for AI safety is inadequate to address critical questions. Only a sociotechnical approach can truly limit current and potential dangers of advanced AI.

Why safety? One could view the shift to safety with cynicism. Big Tech, weary from bad publicity, is seizing the chance to be viewed as saviors from algorithmic harms, not perpetrators of them. Sociotechnical approaches recognize and reject "safety-washing"—giving lip service to safe AI systems, without requisite commitments and practices to ensure this is the case—and call for transparency and accountability to keep companies honest.

"We are making a familiar error."

What does it mean to make AI systems safe, and what values and approaches must be applied to do so? Is it about "alignment," ensuring that deployment of AI complies with some designers' intent? Or is it solely about preventing the destruction of humanity by advanced AI? These goals are clearly insufficient. An AI system capable of annihilating humankind, even if we managed to prevent it from doing so, would still be among the most powerful technologies ever created and would need to abide by a much richer set of values and intentions. And long before such powerful "rogue" AI systems are built, many others will be made that people will use dangerously in their self-interest. Years of sociotechnical research show that advanced digital technologies, left unchecked, are used to pursue power and profit at the expense of human rights, social justice, and democracy. Making advanced AI safe means understanding and mitigating risks to those values, too. And a sociotechnical approach emphasizes that no group of experts (especially not technologists alone) should unilaterally decide what risks count, what harms matter, and to which values safe AI should be aligned. Making AI safe will require urgent public debate on all of these questions and on whether we should be trying to build so-called "god-like" AI systems at all.

How should AI systems be made safe? Narrowly technical approaches are not enough. Instead, AI's use must be targeted in the context of broader sociotechnical systems in which it is always embedded. This means considering the political economy of AI and recognizing when over-indexing on hypothetical future harms predictably risks entrenching the power of a few leading companies, leaving major current challenges unaddressed. It means emphasizing the scarce environmental resources, expropriated data, and exploited labor used to make advanced AI systems. It means considering how advanced AI will be used—for example, to further authoritarianism and illiberalism through mass surveillance and manipulation—which necessitates understanding people and societies, not just their technologies. And it means prioritizing empirically grounded work on actual AI systems and the risks they pose, rather than the (perhaps impossible) task of designing technical mitigations for risks from systems that do not yet exist.

Safety on whose terms? The field of narrow technical AI safety lacks ideological and demographic diversity; it is a near-monoculture and therefore inadequate to the intellectual breadth and rigor required of its mission. Its practitioners, disproportionately white, male, and advantaged, are often drawn from the Silicon Valley social movements of "Effective Altruism" and "Rationalism." A sociotechnical approach resists these structural imbalances, creating spaces for the wider participation necessary to steer our technological future on the basis of equal concern and common humanity. As well as a moral imperative, this is essential for understanding AI risks and building effective solution sets. Women and people of color researchers and advocates have led the way in both revealing the harms of new technologies and in mitigating them; their exclusion from AI safety delegitimizes that field. Sociotechnical AI safety shifts power away from the monoculture.

We are making a familiar error. Faced with disorienting technological change, people instinctively turn to technologists for solutions. But the impacts of advanced AI cannot be mitigated through technical means alone; solutions that do not include broader societal insight will only compound AI's dangers. To really be safe, society needs a sociotechnical approach to AI safety.

–Seth Lazar and Alondra Nelson

10.1126/science.adi8982

AI safety on whose terms?

Seth Lazar and Alondra Nelson

Science, **381** (6654), .
DOI: 10.1126/science.adi8982