AI and Catastrophic Risk | Journal of Democracy https://www.journalofdemocracy.org/ai-and-catastrophic-risk/

AI and Catastrophic Risk

SHARE

AI with superhuman abilities could emerge within the next few years. We must act now to protect democracy, human rights, and our very existence.

By Yoshua Bengio

September 2023

How should we think about the advent of formidable and even superhuman artificial intelligence (AI) systems? Should we embrace them for their potential to enhance and improve our lives or fear them for their potential to disempower and possibly even drive humanity to extinction? In 2023, these once-marginal questions captured the attention of media, governments, and everyday citizens after OpenAI released ChatGPT and then GPT-4, stirring a whirlwind of controversy and leading the Future of Life Institute to publish an open letter in March. That letter, which I cosigned along with numerous experts in the field of AI, called for a temporary halt in the development of even more potent AI systems to allow more time for scrutinizing the risks that they could pose to democracy and humanity, and to establish regulatory measures for ensuring the safe development and deployment of such systems. Two months later, Geoffrey Hinton and I, who together with Yann Le Cun won the 2018 Turing Award for our seminal contributions to deep learning, joined CEOs of AI labs, top scientists, and many others to endorse a succinct declaration: "Mitigating the risk of extinction from

2 von 30 05.11.23, 21:39

AI and Catastrophic Risk | Journal of Democracy https://www.journalofdemocracy.org/ai-and-catastrophic-risk/

AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war." Le Cun, who works for Meta, publicly disagreed with these statements.

This disagreement reflects a spectrum of views among AI researchers about the potential dangers of advanced AI. What should we make of the diverging opinions? At a minimum, they signal great uncertainty. Given the high stakes, this is reason enough for ramping up research to better understand the possible risks. And while experts and stakeholders often talk about these risks in terms of trajectories, probabilities, and the potential impact on society, we must also consider some of the underlying motivations at play, such as the commercial interests of industry and the psychological challenge for researchers like myself of accepting that their research, historically seen by many as positive for humankind, might actually cause severe social harm.

There are serious risks that could come from the construction of increasingly powerful AI systems. And progress is speeding up, in part because developing these transformative technologies is lucrative—a veritable gold rush that could amount to quadrillions of dollars (see the calculation in Stuart Russell's book on pp. 98–99). Since deep learning transitioned from a purely academic endeavor to one that also has strong commercial interests about a decade ago, questions have arisen about the ethics and societal implications of AI—in particular, who develops it, for whom and for what purposes, and with what potential consequences? These concerns led to the development in 2017 of the Montreal Declaration for the Responsible Development of AI and the drafting of the Asilomar AI Principles, both of which I was involved with, followed by many more, including the OECD AI Principles (2019) and the UNESCO

4 von 30 05.11.23, 21:39

AI and Catastrophic Risk | Journal of Democracy https://www.journalofdemocracy.org/ai-and-catastrophic-risk/

Recommendation on the Ethics of Artificial Intelligence (2021).

Modern AI systems are trained to perform tasks in a way that is consistent with observed data. Because those data will often reflect social biases, these systems themselves may discriminate against already marginalized or disempowered groups. The awareness of such issues has created subfields of research (for example, AI fairness and AI ethics) as well as the development of machine-learning methods to mitigate such problems. But these issues are far from being resolved as there is little representation of discriminated-against groups among the AI researchers and tech companies developing AI systems and currently no regulatory framework to better protect human rights. Another concern that is highly relevant to democracy is the serious possibility that, in the absence of regulation, power and wealth will concentrate in the hands of a few individuals, companies, and countries due to the growing power of AI tools. Such concentration could come at the expense of workers, consumers, market efficiency, and global safety, and would involve the use of personal data that people freely hand over on the internet without necessarily understanding the implications of doing so. In the extreme, a few individuals controlling superhuman AIs would accrue a level of power never before seen in human history, a blatant contradiction with the very principle of democracy and a major threat to it.

The development of and broad access to very large language models such as ChatGPT have raised serious concerns among researchers and society as a whole about the possible social impacts of AI. The question of whether and how AI systems can be controlled, that is, guaranteed to act as intended, has been asked for many years, yet there is still no satisfactory answer—although the AI safety-

6 von 30 05.11.23, 21:39

AI and Catastrophic Risk | Journal of Democracy https://www.journalofdemocracy.org/ai-and-catastrophic-risk/

research community is currently studying proposals.

Misalignment and Categories of Harm

To understand the dangers associated with AI systems, especially the more powerful ones, it is useful to first explain the concept of *misalignment*. If Party A relies on Party B to achieve an objective, can Party A concisely express its expectations of Party B? Unfortunately, the answer is generally "no," given the manifold circumstances in which Party A might want to dictate Party B's behavior. This situation has been well studied in both economics, in contract theory, where Parties A and B could be corporations, and in the field of AI, where Party A represents a human and Party B, an AI system. If Party B is highly competent, it might fulfill Party A's instructions, adhering strictly to "the letter of the law," contract, or instructions, but still leave Party A unsatisfied by violating "the spirit of the law" or finding a loophole in the contract. This disparity between Party A's intention and what Party B is optimizing is referred to as a *misalignment*.

One way to categorize AI-driven harms is by considering intentionality —whether human operators are intentionally or unintentionally causing harm with AI—and the kind of misalignment involved: 1) AI used intentionally as a powerful and destructive tool—for instance, to exploit markets, generate massive frauds, influence elections through social media, design cyberattacks, or launch bioweapons—illustrating a misalignment between the malicious human operator and society; 2) AI used unintentionally as a harmful tool—for instance, systems that discriminate against women or people of color or systems that inadvertently generate political polarization —demonstrating a misalignment between the human operator and the AI; and 3) loss of control of an AI

8 von 30 05.11.23, 21:39