https://doi.org/10.1038/s42256-024-00967-8

Strategies needed to counter potential AI misuse

Check for updates

Researchers urgently need more guidance to help them identify and mitigate potential risks when designing projects that involve AI developments.

Since the dawn of civilization, society grappled with the challenge that newly invented technology, while bringing benefits and ensuring progress, also carries the potential for misuse and harmful applications. Given the fast-rising capabilities and wide deployment of artificial intelligence (AI) tools, it has become even more urgent to examine the negative side of innovation. Advances in AI have led to transformative advances in scientific fields, such as the capability to predict and generate protein structures, which was recently awarded the Nobel Prize in Chemistry. However, even as AI accelerates positive scientific advances, the scale for potential misuse is also growing, and policies and guidelines need to catch up.

For example, a Correspondence in this issue considers how AI can speed up genetic and genomic research, including gene-editing experiments that might allow the resurrection of extinct species or the creation of new dangerous pathogens. The authors highlight the need to think through moral questions and unwanted applications sooner rather than later.

The piece is among a growing number of publications and initiatives that highlight concerns about the misuse and dual use of AI technology in specific fields. Others have shown that AI models intended for drug development with specific disease targets could, in the wrong hands, be deployed to generate new toxic compounds[1]. Similarly, AI-driven materials design methods could lead to the production of environmentally harmful substances[2]. In both cases, the authors expressed shock and surprise about the scope for harmful applications once they were prompted to examine them[3].

With slow developments in regulation, AI researchers face uncertainty and need guidance to navigate the landscape of ethical and misuse concerns. A Perspective in this issue by Trotsyuk et al. proposes a framework to help researchers identify and mitigate negative uses of AI tools in biomedical applications. The authors have backgrounds ranging from bioengineering, medicine and drug discovery to AI ethics, and describe guidelines to provide support for researchers in navigating questions on harmful downstream implications and misuse at the start of a project involving AI development.

A first step in this framework is for researchers to assess the possible risks and benefits of their project at an early stage, among others by engaging with ethicists and affected groups. If these efforts reveal a sizable risk of harm or misuse, researchers can turn to various tools that can help to mitigate risk, which the authors have categorized into three types: implementing existing ethical frameworks and regulatory measures; using off-the-shelf tools and strategies (such as adversarial testing); and seeking design-specific solutions (such as those related to data collection and management).

The authors built the framework after examining three case studies of AI applications in biomedicine, namely drug discovery, generative models for synthetic data, and ambient intelligence. For the latter, AI-enabled Internet of Things (IoT) devices could be used with great benefit for patient monitoring and in the care of older people. However, reviewing the potential ethical and societal impact shows that the technology can also negatively affect privacy, potentially by leaking health information or leading to intrusive surveillance. A next step is to apply mitigating strategies such as introducing disclosure and opt-out systems, adversarial testing, and privacy-enhancing measures. If, after further assessment, the risks are likely to outweigh benefits, researchers should reconsider the project.

The study is US-oriented, a limitation noted in the article, and could be broadened to more geographically diverse case studies. But initiatives such as these are badly needed to support researchers in tackling concerns at an early project development stage, and to encourage a 'safety by design' mindset rather than 'moving fast and breaking things'. Efforts to develop guidance for researchers on AI safety and misuse concerns must be ongoing to keep up with the fast pace of AI developments.

Published online: 18 December 2024

References

1. Urbina, F., Lentzos, F., Invernizzi, C. & Ekins, S. *Nat. Mach. Intell.* **4**, 189–191 (2022).
2. Shankar, S. & Zare, R. N. *Nat. Mach. Intell.* **4**, 314–315 (2022).
3. [Editorial]. *Nat. Mach. Intell.* **4**, 313 (2022).