Editorial

https://doi.org/10.1038/s42256-023-00782-7

A year of racing ahead with AI and not breaking things

Looking back at a year of escalating, divisive debates in AI safety and who determines the agenda.

One of the most debated topics in artificial intelligence (AI) in 2023 has been safety concerns, as large AI models are rapidly becoming more powerful and widely available. Concerns about safety, alignment, ethical risks and the societal and environmental effects of AI have been active discussion points for several years, but the public agenda has seen a shift in emphasis as discussions about existential risk due to out-of-control AI have become more prominent. These concerns have become mainstream in the past year partly because AI leaders such as Yoshua Bengio and Geoffrey Hinton have expressed concern about near-future scenarios in which powerful AI models are out of control, either due to malicious use by terrorists or rogue states, or because they acquire superhuman intelligence, become self-aware and autonomously pursue their own goals.

AI systems have seen rapid development in the past decade since the rise of deep learning around 2012, but even swifter progress has been made with large language models and generative AI in the last few years. With the release of the generative AI model ChatGPT around a year ago, the world woke up to the power of such models, and the likely disruptive effect that they will have on many areas of society. There are also serious dual-use concerns – malicious actors can use generative AI to scale up harmful and criminal schemes such as in misinformation and cybercrime.

Those paying attention may have been surprised to find out that OpenAI, the company behind ChatGPT, is pursuing as its core mission the development of artificial general intelligence (AGI), as described on their website. The company has laid out a plan to deploy increasingly powerful AI in stages, giving everyone "incredible new capabilities" and benefitting all of humanity, while addressing potential risks. OpenAI asks us to trust them to get it right with their approach of gradual, safe deployment. The release of ChatGPT has been a large-scale experiment in this direction, in which millions of users provide free feedback (in addition to paid annotators). The model is adjusted, and anti-toxicity filters are added when concerns are noticed. However, how to safely, fully align a model such as ChatGPT with human intention is unknown.

OpenAI is not the only player in frontier AI models (or to pursue AGI). Soon after ChatGPT was released, other companies joined the fray with generative models, with or without chatbot interfaces, such as LLaMA by Meta, ERNIE by Baidu, Claude by Anthropic, and most recently Gemini by Google. GPT-4, an impressive multimodal version of the model that underlies ChatGPT, was released in March 2023. In a recent survey by researchers at Google DeepMind on progress towards AGI[1], several of these models are classified as 'emerging' AGI, the lowest of five levels, which means that the models are at least as good as, or slightly better than, a human who was not specifically trained for the specific task.

With this whirlwind of AI developments in early 2023, in March several key figures in AI research and industry signed an open letter from the Future of Life Institute calling for a 6-month pause on 'giant AI experiments'. The letter recommends that powerful AI systems should be developed only once we are confident that their effects will be positive and their risks will be manageable. However, this initiative received considerable backlash as experts with research backgrounds in AI ethics argued that the open letter is too narrowly focussed on unknown, sci-fi risks, while there are many current, known risks that require equal, if not greater attention.

There were similar concerns in response to the announcement of the AI safety summit, which was organized by the UK on 1–2 November with representatives from 28 countries and several industries and civil societies. The effort to develop guardrails for AI development, and to start international collaboration to ensure safe development of AI, is surely welcome. But the summit has been criticized for its focus on long-term risks and the reliance on self-regulation by the industry. Another open letter published by several civil societies at the start of the summit calls for regulatory action to address the full range of risks that AI systems can raise, including current risks that are already affecting human rights.

Just before the safety summit, on 30 October, US President Joseph Biden issued an executive order on 'New Standards for AI Safety and Security'. It was the first executive order on AI by the USA, and provides guidance for federal agencies, along with standards and definitions in AI. Among its actions, it requires that developers of the most powerful AI systems share their safety test results and other crucial information with the US government, and calls for the development of standards, tools and tests to help to ensure that AI systems are safe, secure and trustworthy[2].

Such requirements are in line with recommendations formulated by a group of industry and academic experts in AI safety, from OpenAI, Anthropic and Google DeepMind, among others. The researchers call for steps in requiring registration and reporting of powerful 'frontier' AI models, risk-assessments and post-deployment monitoring, relying both on regulation and self-regulation[3]. They discuss emerging risks and the need to prepare for unpredictable developments of dangerous capabilities in AI models. However, it should be noted that there have already been predictable and clear harms from AI applications happening for several years – such as algorithmic discrimination, exploitation of low-paid data annotators, copyright infringement and deepfake attacks, and these have also not been addressed. As 2024 is around the corner and the AI safety debate will continue, we support calls to ensure that the agenda is not just determined by representatives from big tech companies, and that national and international regulation take into account a wider, more global range of voices to address realistic AI harms, future and present.

Published online: 18 December 2023

References

1. Morris, M. R. et al. Preprint at https://doi.org/10.48550/arXiv.2311.02462 (2023).
2. Jones, N. *Nature* **623**, 229–230 (2023).
3. Anderljung, M. et al. Preprint at https://doi.org/10.48550/arXiv.2307.03718 (1012).