https://doi.org/10.1038/s42256-023-00755-w

AI reality check

AI-generated media are on the rise and are here to stay. Regulation is urgently needed, but in the meantime creators, users and content distributors need to pursue various ways, and adopt various tools, for responsible generation, sharing and detection of AI-generated content.

The fast rise of generative AI tools has made a large impact on the world, providing myriad new ways to create content, such as images, audio, video and text, and with applications in entertainment, education, industry, scientific research and communication and beyond. The pace of developments is high as new tools and improved versions of existing tools are constantly released. To name a few recent developments among many, the latest version of chatbot Ernie was launched on the Chinese platform Baidu[1], rivalling ChatGPT capabilities, and OpenAI's prompt-to-image generating tool DALL-E 3 was released and integrated in Bing Chat[2].

Unfortunately, there are also ample opportunities for malicious use of generative AI, as documented in several news stories this year. There is growing worry that a massive spread of deepfakes and disinformation could undermine democratic elections in 2024 in countries around the world, if only as individual voters lose confidence in any news and withdraw further in filter bubbles[3]. Another concern is how generative AI has recently been used to enable voice cloning-based impersonation in the form of vishing (voice phishing) attacks and phone call imposter scams[4]. In science, research communities are concerned about the flood of made-up research papers and fake results, raising the possibility of a breakdown of trust in scientific literature[5].

Technical solutions are being pursued to save reality from getting drowned out by fake and harmful information. One line of attack is to build, or re-build, trust by ensuring that synthetic content is rigorously identified or labelled — for example, with disclaimers preceding an article to warn a reader that the text has been AI-generated. Synthetic images can be watermarked with signals added to pixels that are imperceptible to viewers, but that can be picked up by image classification tools. For example, a watermarking application was recently demonstrated by Google DeepMind; the tool, SynthID, is designed to be particularly resistant to tampering, but although more robust than previous attempts, the technology might be circumvented in the future and has not yet been made widely available[6].

A complementary approach is to develop a standard for adding cryptographic tags to digital files to incorporate information on the origin of content, or its 'provenance', a route that is being promoted by the Coalition for Content Provenance and Authenticity (C2PA). The organization was founded in 2021 by Microsoft and Adobe, and brings together industry efforts in content authentication. In a Correspondence in this issue, Rieger et al. point out that a possible downside of this approach is that editing a document invalidates its associated digital signature. Addressing the issue by requiring authentication of content creators and editors is prone to privacy risks, especially when the identity of those involved in the modification of the files needs to be protected. The authors highlight the possibility of enhancing cryptographic methods with mathematical tools known as 'zero-knowledge proofs', which can safely trace any modifications without revealing sensitive information.

But despite best efforts to encourage tagging synthetic content, methods are also needed to detect undeclared, potentially malicious synthetic content. Various forensic tools exist that can pick up statistical features or tell-tale signs in AI-generated content. Of course, unavoidably, the generative AI models involved get better as they learn to avoid detection[7]. As Menczer et al. write in a recent Correspondence, researchers need to pursue many avenues to develop strategies, including AI-based ones, to keep up with the capabilities of generative AI tools and the creativity of those deploying them in harmful schemes. For instance, beyond detecting standalone AI-generated content, methods can be aimed at detecting suspicious behaviour at the group level, unveiling botnet activity[8].

There is a growing role for content distributors and publishers in curbing the spread of harmful generated content. Earlier this year, the Partnership on AI (PAI, a non-profit community of industry and other organizations) published a report on responsible practices for synthetic media with recommendations for different stakeholders: those building generative AI technology, creators of synthetic content, and distributors and publishers of such content. The PAI calls on publishers and distributors to provide policies outlining their approach to deal with synthetic content. The reality is that it is challenging to keep up with the quickly developing capabilities of generative AI tools and applications, and such policies will need to be regularly updated as the world is racing ahead with inventing generative AI tools and applications.

Ultimately, regulation at the national and international level is necessary to ensure that technology companies take responsibility and prevent their AI tools from harming society[9]. In the meantime, content users, creators, platforms, distributors and publishers need to adopt an arsenal of approaches to detect, label and safely share AI-generated content.

Published online: 19 October 2023

References

1. David, E. *TheVerge* https://go.nature.com/3PNI9eN (2023).
2. Diaz, M. *ZDNET* https://go.nature.com/3Q5PO9p (2023).
3. *The Economist* https://go.nature.com/3Q3Fi2C (2023).
4. Upton-Clark, E. *Insider* https://go.nature.com/3RNZlTP (2023).
5. Jones, N. *Nature* **621**, 6656–679 (2023).
6. Pierce, D. *The Verge* https://go.nature.com/46vpNpG (2023).
7. Boneh, D., Grotto, A. J., McDaniel, P. & Papernot, N. *HAI Policy Briefs* https://go.nature.com/3F6o2n2 (Stanford Univ., 2020).
8. *Wired* https://go.nature.com/3ZI50N6 (2023).
9. European Parliament. *EU AI Act* https://go.nature.com/3RImQOd (2023).