https://doi.org/10.1038/s42256-022-00598-x

# Much to discuss in AI ethics



**2022 has seen eye-catching developments in AI applications. Work is needed to ensure that ethical reflection and responsible publication practices are keeping pace.**

So-called large language models (LLMs), which are based on the deep neural network transformer architecture and pretrained with large amounts of unlabelled data, have led to breakthrough capabilities in the past few years[1]. Models such as GPT-3 from OpenAI and LaMDA from Google, with many billions of parameters, can perform a range of language tasks, from translation to conversational agents and text generation. Most recently, OpenAI released ChatGPT, a chatbot that is taking the world by storm with its impressive abilities to give interesting and engaging responses. Moreover, LLMs can be developed into a range of other generative tools when trained on specific types of data from the web: for example, in code writing, such as Copilot from Microsoft, which is trained on GitHub repositories, and in image generation[2], such as DALL-E (OpenAI) or Stable Diffusion. Recently, Meta announced a new tool called Galactica[3], which is trained on preprints and free-to-access papers, for summarizing, annotating and gathering scientific knowledge. An online demo version of Galactica was withdrawn, however, within days after much online criticism about the potential for harmful outputs[4]. For instance, it was shown that the tool could be used to write authoritative-sounding scientific papers or explanations that were incorrect and potentially dangerously so.

A persistent problem with many experimental AI tools, such as those based on LLMs, is that they have many limitations that are not sufficiently understood, but that could lead, intentionally or unintentionally, to harmful applications. Those who contribute to AI developments, therefore, need to engage more with ethical processes to ensure responsible publication and release of AI tools[5–7]. This is urgent and necessary given the reach of AI, with many applications being pervasive in society and posing a substantial risk of potential harm and misuse. The AI community should look towards the biomedical sciences that lead the way in discussing emerging research topics where risk of harm and ethical concerns are identified, and where consensus is reached over rules and guidance in research outputs, such as in gene editing and stem cell research.

As highlighted in a Comment in this issue by Srikumar et al., several AI venues now have ethical processes in place, including NeurIPS, the biggest annual AI conference. NeurIPS introduced a paper submission checklist in 2020 and added an ethical review process for selected papers in 2021. The checklist prompts authors to discuss ethical considerations such as potential bias in training data and also to disclose potential negative applications of their work in a broader impact statement. A list of specific examples are provided, including use in harmful types of surveillance (for instance, in research on facial imaging) and damage to the environment (for instance, in applications that help with fossil fuel exploration). *Nature Machine Intelligence* currently requests authors to provide an ethical and societal impact statement in papers that involve identification or detection of humans or groups of humans, including behavioural and socio-economic data. An example can be found in an Article in this issue, on a deep learning approach in lip reading for multiple languages that has a section on ethical considerations.

The discussions around Galactica, and also ChatGPT[8], echo concerns about AI tools and LLMs that have been repeatedly voiced over the past few years[9]. In particular, critics warn that such models make up answers, without any real understanding or concern about being correct or not. Moreover, the output is based on statistical information in large amounts of data that usually have not been curated or checked and that include problematic biases. Producing realistic but often incorrect and potentially biased output at scale seems a perilous capability — fuelling concerns regarding the use of AI in harmful deepfakes and the spread of misinformation. Naturally, these are not the intended uses of AI systems such as ChatGPT, Galactica and other tools. In fact, an often-heard response to defend AI applications is that technology itself is neutral. Like many in the community, we beg to differ. Technology is almost never neutral: at all steps in research and development, humans and human data are involved with certain backgrounds, goals and biases. It is natural to request that researchers engage in ethical reflection, enlisting advice from ethics, legal and other experts, when developing and releasing a tool. Particular attention is needed for applications that involve large amounts of human-generated data, such as data sourced from the web.

Srikumar et al. recommend that research is carried out to study current ethical processes at AI conferences and journals, and to examine their effectiveness. They argue that there should be more experiments with initiatives in ethics and responsible publication. Moreover, they emphasize the need for venues where public debate and discussion can take place. Various dedicated workshops and conferences in the area of AI ethics already exist, but what is needed are platforms where concrete steps and consensus about next steps and guidelines are made. This seems especially important with regard to deciding what constitutes serious dual use risk, where malicious use of an AI tool can lead to weaponization and other dangerous applications. The topic was highlighted earlier this year in a Comment[10] where researchers raised the alarm over generative approaches in drug discovery; they demonstrated with a simple thought experiment the ease of using such a tool to (computationally) optimize for toxicity rather than for beneficial health effects. Community discussion is necessary to discuss how to prevent such misuse and, for instance, to decide on whether controlled release of models and data is required.

We look forward to important discussions — and decisions — in ethics and responsible publication in 2023.

Published online: 19 December 2022

## References

1. *Nat. Mach. Intell.* **3**, 737 (2021).
2. *Nat. Mach. Intell.* **4**, 733 (2022).
3. Taylor, R. et al. Preprint at *arXiv* arxiv.org/abs/2211.09085v1 (2022).
4. Heaven, W. D. Why Meta's latest large language model survived only three days online. *MIT Technology Review* https://www.technologyreview.com/2022/11/18/1063487/meta-large-language-model-ai-only-survived-three-days-gpt-3-science/ (2022).

5. Prunkl, C. E. A. et al. *Nat. Mach. Intell.* **3**, 104–110 (2021).
6. Hecht, B. et al. Preprint at *arXiv* https://arxiv.org/ftp/arxiv/papers/2112/2112.09544.pdf (2018).
7. PAI Staff. Managing the risks of AI research: six recommendations for responsible publication 247. *Partnership on AI* https://partnershiponai.org/paper/responsible-publication-recommendations/ (2021).
8. Knight, W. ChatGPT's Most Charming Trick Is Also Its Biggest Flaw. *Wired* https://www.wired.com/story/openai-chatgpts-most-charming-trick-hides-its-biggest-flaw/ (2022).
9. Bender E. M. et al. in Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, FAccT '21. 610–23 (*Association for Computing Machinery*, 2021).
10. Urbina, F., Lentzos, F., Invernizzi, C. & Ekins, S. *Nat. Mach. Intell.* **4**, 189–191 (2022).