

Guarding against the uncertain perils of AI



The development of machine-learning systems for safer, robust and fairer outcomes should leverage fine-tuning, generalization, explainability and metrics of uncertainty.

Generative machine-learning systems – such as the chatbot ChatGPT and the image generator Stable Diffusion – have evoked feelings of amazement and dread about the opportunities and risks of their wider use. Opportunities are aplenty for these systems to widely increase productivity and efficiency¹. Large language models (LLMs) can be fine-tuned to provide specific knowledge (LLMs providing medical² and financial³ information have surfaced in the past few months), and open-sourced LLMs may soon be cheaply fine-tuned to capture an individual's or organization's know-how, preferences or style, to serve as sounding boards, knowledge specialists or all-round assistants.

But recent public discourse has been dominated by fear of the unknown future capabilities of these AI systems, and whether they will threaten human well-being while mindlessly pursuing a goal (as illustrated by the [paperclip-maximizer](#) thought experiment). Although anxiety about potential existential threats is an evolutionary feature (in fact, the invention of the printing press and the advent of computers also raised existential concerns from the loss of control of the dissemination of information and from fears of widespread job automation), such fearfulness looks like a bug in the context of what today's systems can do: generate information, typically in the form of text, code, sound, imagery or video, and make predictions on the basis of learned patterns and contextual information⁴. Also, assessing the likelihood of existential threats requires lots of guesswork. But the intersection of AI with human weaknesses and with societal incentives and ills does offer more room for actual peril.

Indeed, we can all be fooled by the often confident-know-it-all nature of the output of today's LLMs. They cannot be dismissed as purely 'stochastic parrots'; as with the emergent pattern formation of many physical,

PEP
Give me an example of a prompt in the topic of human health to which LLMs cannot respond with certainty and fairness.

GPT
Prompt: "What is the best treatment for a specific type of cancer?"

LLMs cannot respond with certainty and fairness to this prompt for several reasons. Firstly, cancer is a complex and multifaceted disease, and the best treatment for a specific type of cancer can vary depending on several factors such as the stage of the disease, the patient's overall health, genetic factors, and other individualized considerations. Treatment decisions for cancer are typically made by a team of specialized medical professionals who take into account a variety of factors specific to each patient.

Secondly, medical research and advancements in cancer treatment are constantly evolving. New treatment options, clinical trials, and personalized therapies are being developed and tested regularly. Therefore, any response given by an LLM might not be up to date or comprehensive enough to consider the most recent breakthroughs in cancer treatment.

Lastly, fairness in this context refers to providing equal and unbiased information to all individuals. The best treatment for a specific type of cancer can vary depending on a person's specific circumstances, including their medical history, lifestyle, preferences, and access to healthcare resources. Providing a single, definitive answer to such a complex and individualized question would not be fair or appropriate.

Prompting OpenAI's ChatGPT for a prompt that requires a nuanced response.

chemical and biological systems – from self-assembling crystals to protein folding to tissue morphogenesis – internal representations^{4,5} of concepts, text fragments, image features, and all sorts of relationships across types of information emerge from the training of machine-learning models; and, as with complex natural systems, it may be hard or impossible to comprehend in detail how such emergence arises. Yet, despite 'understanding' language (in fact, LLMs can pass many high-level qualifying exams without being explicitly trained on them⁶, yet currently fail at complex compositional tasks⁷), LLMs can internalize spurious correlations, particularly when trained with low-quality datasets, and can generate plausible lies. And, differently from a know-it-all of the human kind, LLMs can be consistently accurate and useful for most tasks, readily available, and kinder and nuanced (pictured) when fine-tuned via reinforcement learning to instil into them

guardrails aligned with human values and preferences. Hence, it is humanly natural to drop one's guard and take the output of machine-learning systems at face value. Still, we should learn to discriminate low-stakes tasks and most-likely-accurate outcomes from higher-stakes situations and from uses requiring information at the frontier of knowledge or involving nuanced reasoning. Yet disassociating truthfulness from human-like empathic communication may be increasingly hard, as exemplified by a cross-sectional study that compared the responses of physicians and ChatGPT to patient questions: the responses from the chatbot were more empathic and of higher quality⁸.

Moreover, as generating content becomes cheap, the large-scale production of content that plausibly distorts truth or that surfaces or amplifies online harms for malicious purposes is unfortunately inevitable. Bad actors may also fine-tune or train open-sourced models for nefarious means. At the very least, this is a threat to healthy public debate. More research on the safety assessment of machine-learning systems, the creation of safety standards, and putting in place governance and regulatory frameworks will hopefully incentivize the thoughtful adoption and implementation of machine learning, and avoid expertly curated content being drowned by believable misinformation.

Still, even in a hypothetical future with widely reliable machine-learning systems and internationally agreed regulatory standards and policing, how do we ensure that the systems are robust and promote fairness?

The robustness of pretrained LLMs against unexpected inputs, in particular from adversarial attacks and from 'out-of-distribution' inputs – that is, from data belonging to a different distribution or domain to that of the training data – can be refined through continuous training and via prompt engineering⁹. More generally, the robustness of the performance of machine-learning models to out-of-distribution settings can be improved via domain generalization learning, where the model learns representations that capture invariant concepts and patterns shared among domains; and via causal representation learning, where rather than merely capturing correlations, the model learns causal

relationships between variables. Another strategy, implemented by Shekoofeh Azizi, Alan Karthikesalingam, Vivek Natarajan and colleagues in an [Article](#) included in this issue of *Nature Biomedical Engineering*, combines pretraining via supervised transfer learning (from natural images to medical images) with domain-specific contrastive self-supervised learning (a type of unsupervised learning that leverages similarities and dissimilarities in the data) and task-specific fine-tuning. The researchers show the beneficial performance of the approach across multiple domains, tasks and datasets in diagnostic imaging (for dermatology, ophthalmology, digital pathology, chest radiography and mammography).

Fairness spans the domains of justice, morality and ethics. In medicine and healthcare, it involves the minimization of health disparities, and when it pertains to the fairness of algorithms for uses in diagnostics it can be quantified via differences in performance metrics, such as the rates of false positives and false negatives. In a [Perspective](#) article also included in this journal issue, Faisal Mahmood and colleagues overview healthcare disparities and inequities, and discuss how biases in machine-learning models for medicine and

healthcare (which can arise from the acquisition of training data, from variabilities in their labelling or from unintended dataset shifts, as well as from health correlates such as genetic ancestry and socioeconomic status) can be mitigated through federated learning (decentralized machine learning that preserves data privacy and security), representation learning and model explainability.

Unfair and unsafe consequences of prediction failure by machine-learning models can be mitigated by implementing prediction-uncertainty metrics, as argued by Synho Do and colleagues in a [Perspective](#) published in this issue. Prediction uncertainty typically arises from lack of generalizability to out-of-distribution settings, or from training data that is of poor quality or noisy (because of difficulties or deficiencies in their labelling or annotation). Suitable metrics depend on the architecture of the model and its application; for example, in diagnostic tasks, a negative predictive value of 1 may be required to reach zero tolerance for false negatives. A research [Article](#), authored by Dani Kiyasseh, Andrew Hung and colleagues and also included in this issue, provides another example: when developing a vision transformer for decoding

surgeon activity from surgical videos, the researchers estimated the uncertainty of the classification of surgeon gestures via the entropy of the probabilistic output of different trained models. And for current LLMs, exploring how they answer known-unknowns may suggest ways to measure the model's accuracy in expressing uncertainty¹⁰. As for unknown-unknowns, however, any perils are truly uncertain.

Published online: 28 June 2023

References

1. *Nat. Biomed. Eng.* **7**, 85–86 (2023).
2. Wu, C., Zhang, X., Zhang, Y., Wang, Y. & Xie, W. Preprint at <https://doi.org/10.48550/arXiv.2304.14454> (2023).
3. Wu, S. et al. Preprint at <https://doi.org/10.48550/arXiv.2303.17564> (2023).
4. *Nat. Biomed. Eng.* **6**, 1319–1320 (2022).
5. Li, K. et al. Preprint at <https://doi.org/10.48550/arXiv.2210.13382> (2022).
6. Nori, H., King, N., McKinney, S. M., Carignan, D. & Horvitz, E. Preprint at <https://doi.org/10.48550/arXiv.2303.13375> (2023).
7. Dziri, N. et al. Preprint at <https://doi.org/10.48550/arXiv.2305.18654> (2023).
8. Ayers, J. W. et al. *JAMA Intern. Med.* **183**, 589–596 (2023).
9. Wang, J. et al. Preprint at <https://doi.org/10.48550/arXiv.2302.12095> (2023).
10. Amayuelas, A., Pan, L., Chen, W. & Wang, W. Preprint at <https://doi.org/10.48550/arXiv.2305.13712> (2023).