College of Medicine and Public Health, Flinders University, Adelaide SA 5042, Australia

Correspondence to: A M Hopkins
ashley.hopkins@flinders.edu.au
https://orcid.org/0000-0001-7652-4378

Cite this as: *BMJ* 2024;384:q596
http://dx.doi.org/10.1136/bmj.q596

FAST FACTS

Quality and safety of artificial intelligence generated health information

Generative artificial intelligence (AI) is advancing rapidly and has the potential to greatly improve many aspects of society, including health. The risks of potentially harmful consequences, however, necessitate effective oversight and mitigation measures. This article highlights distinct forms of health related risks of generative AI, with corresponding options for mitigating risk.

Michael J Sorich, Bradley D Menz, Ashley M Hopkins

Although artificial intelligence (AI) holds considerable promise for positive effects on society, it also has the potential for harmful consequences, which may occur either unintentionally or because of misuse. Applications, such as ChatGPT, Gemini, Midjourney, and Sora, showcase generative AI's capability to create high quality text, audio, video, and image content. The rapid advancement of AI technologies requires an equally rapid escalation of efforts to identify and mitigate risks. New disciplines, such as AI Safety and Ethical AI, broadly aim to ensure that current and future AI operates in a manner that is safe and ethical.

This article focuses on generative AI—a technology with substantial potential to transform how communities seek, access, and communicate information, including about health. Table 1 outlines a glossary of key terms used in the article. Given that more than 70% of people turn to the internet as their first source of health information,[1] it is crucial to identify common types of risks associated with AI technologies and to introduce effective vigilance structures for mitigating these risks. Notably, as generative AI becomes increasingly sophisticated, it will become more challenging for the public to discern when outputs (text, audio, video) are incorrect. In this article, we aim to differentiate common types of potential risks and highlight emerging ideas for mitigating each type of risk. For simplicity, we often use large language models (LLMs) to illustrate emerging problems, but the concepts and considerations presented apply to generative AI more broadly.

AI errors

Across all types of AI, errors are a common challenge. As the text, audio, and video output of modern generative AI has become increasingly sophisticated, erroneous or misleading responses may be difficult to detect. The phenomenon of "AI hallucination" has gained prominence with the widespread use of AI chatbots (eg, ChatGPT) powered by LLMs. In the health information context, AI hallucinations are particularly concerning because individuals may receive incorrect or misleading health information from LLMs that are presented as fact.[2 3] For members of the general public, who may lack the capability to distinguish between correct and incorrect information, this has considerable potential for harm. For healthcare professionals using LLMs to generate clinical documentation, the generated outputs must be treated as drafts that require careful review for accuracy before finalisation.

Numerous technological strategies are being explored to minimise potential risks associated with generative AI errors. One promising strategy for accurately answering health related questions involves developing generative AI applications that "ground" themselves in relevant sources of information. This approach diverges from earlier methods that relied on responses being generated from model "memory." Instead, many AI applications can now access and subsequently summarise information from up-to-date, authoritative sources. For example, many AI chatbots now incorporate real time internet search capabilities to return responses that summarise and explicitly cite the information source. Another approach is to improve "uncertainty quantification" for generative AI. This involves developing generative AI that better communicates the level of uncertainty associated with its response. Therefore, when unsure about an answer, the response should clearly highlight the uncertainty, thereby allowing the user to interpret the information more appropriately.

Health disinformation

As distinct from AI hallucinations, where incorrect or misleading information is generated inadvertently, it is also possible for malicious actors to intentionally generate incorrect or misleading information using generative AI if effective guardrails are lacking. When incorrect or misleading information is generated deliberately, it is referred to as disinformation. Although disinformation is not new, generative AI may enable the inexpensive creation of diverse, high quality, targeted disinformation at scale.[4 5] This problem is not specific to health, but the effects of enhanced health disinformation are likely to be particularly problematic for society.

One option for preventing AI generated health disinformation involves fine tuning models to align with human values and preferences, including avoiding known harmful or disinformation responses from being generated. An alternative is to build a specialised model (separate from the generative AI model) to detect inappropriate or harmful requests and responses. The specialised model would screen the request before allowing it to be passed to the generative AI model, and the output of the generative AI model would be screened before releasing the

output. In the linked study (doi:10.1136/bmj-2023-078538), we highlighted that many popular AI chatbots and assistants—including ChatGPT, Copilot, Bard, and HuggingChat—lack effective guardrails for preventing the generation of health disinformation.[4]

Initiatives to facilitate the easy identification of AI generated content, such as embedding digital watermarks, are also underway. Progress, however, is still required towards industry standards for identifiable AI generated material. Such efforts would make it easier for content sharing platforms (eg, social media, search engines) to identify and remove inappropriate AI generated content.[6]

Equally critical to countering AI facilitated disinformation is the establishment of robust AI vigilance processes. As generative AI continues to develop, emergent and unforeseen risks are likely to arise, underscoring the importance of ongoing monitoring, fixing identified safeguard vulnerabilities, and transparency. Our study found a lack of transparency among generative AI developers regarding the safeguards and processes implemented to minimise risks from health disinformation, along with a deficiency in responding to and fixing reported vulnerabilities related to health disinformation.[4]

Privacy and bias

The privacy of personal health information must be prioritised during the development and use of generative AI.[7] Private health information should not be used to train generative AI models, as it is difficult to ensure that sensitive information will not leak into model outputs. Healthcare professionals need to also carefully consider the consequences of inputting sensitive patient information into public AI assistants and chatbots for tasks such as drafting clinical summaries, communications, and emails. Generative AI applications often state terms and conditions that allow developers to store and use information entered. The public should also be aware of this to avoid inputting sensitive information. Therefore, for sensitive data, it is important to only use generative AI services that explicitly commit to not retaining data, or to run the generative AI model locally to ensure that health data are not sent to a third party.

Training of generative AI requires vast amounts of text, image, and audio content, often sourced from the internet. In learning from this diverse material, the AI model is at risk of inheriting biases present in the training material, and hence deployment of the AI risks reinforcing existing inequities.[7 8] Despite efforts by developers to mitigate biases, it remains challenging to fully identify and understand the biases of accessible LLMs owing to a lack of transparency about the training data and process.[8] Ultimately, strategies aimed at minimising these risks include exercising greater discretion in the selection of training data, thorough auditing of generative AI outputs, and taking corrective steps to minimise biases identified.

Concluding remarks

A fundamental current challenge is the rapid progress of AI. One consequence of the frequent release of new AI models, or updates to existing AI models, is that performance and associated risks may change rapidly. For example, in our study, Microsoft's Copilot demonstrated effective safeguards preventing the generation of health disinformation in September 2023, but three months later these safeguards were no longer present.[4] Such a finding outlines that frequent ongoing audits of risks and functionalities will be required.

For readers seeking a deeper understanding of AI safety and ethics as it relates to health, we refer to World Health Organization guidance on ethics and governance of AI for health,[9] and the European Parliamentary Research Service report on the applications, risks, ethics, and societal impacts of AI in healthcare.[10] These documents provide valuable insights into responsible deployment and management of AI technologies, emphasising the critical need for ongoing auditing and adaptation in this rapidly evolving specialty.

Funding and competing interests available in the linked paper on bmj.com.

Provenance and peer review: Commissioned; not externally peer reviewed.

Use of generative AI: During the preparation of this work the authors used ChatGPT and Grammarly AI to assist in the formatting and editing of the manuscript to improve the language and readability. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

1 Finney Rutten LJ, Blake KD, Greenberg-Worisek AJ, Allen SV, Moser RP, Hesse BW. Online Health Information Seeking Among US Adults: Measuring Progress Toward a Healthy People 2020 Objective. *Public Health Rep* 2019;134:-25. doi: 10.1177/0033354919874074 pmid: 31513756

2 Hopkins AM, Logan JM, Kichenadasse G, Sorich MJ. Artificial intelligence chatbots will revolutionize how cancer patients access information: ChatGPT represents a paradigm-shift. *JNCI Cancer Spectr* 2023;7:pkad010. doi: 10.1093/jncics/pkad010 pmid: 36808255

3 Lee P, Bubeck S, Petro J. Benefits, Limits, and Risks of GPT-4 as an AI Chatbot for Medicine. *N Engl J Med* 2023;388:-9. doi: 10.1056/NEJMsr2214184 pmid: 36988602

4 Menz BD, Kuderer NM, Bacchi S, etal. Current safeguards, risk mitigation, and transparency measures of large language models against the generation of health disinformation: repeated cross sectional analysis. *BMJ* 2024;384:e078538.

5 Menz BD, Modi ND, Sorich MJ, Hopkins AM. Health Disinformation Use Case Highlighting the Urgent Need for Artificial Intelligence Vigilance: Weapons of Mass Disinformation. *JAMA Intern Med* 2024;184:-6. doi: 10.1001/jamainternmed.2023.5947 pmid: 37955873

6 Hopkins AM, Menz BD, Sorich MJ. Potential of Large Language Models as Tools Against Medical Disinformation-Reply. *JAMA Intern Med* 2024. doi: 10.1001/jamainternmed.2024.0023 pmid: 38407881

7 Meskó B, Topol EJ. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. *NPJ Digit Med* 2023;6:. doi: 10.1038/s41746-023-00873-0 pmid: 37414860

8 Zack T, Lehman E, Suzgun M, etal. Assessing the potential of GPT-4 to perpetuate racial and gender biases in health care: a model evaluation study. *Lancet Digit Health* 2024;6:-22. doi: 10.1016/S2589-7500(23)00225-X pmid: 38123252

9 *Ethics and governance of artificial intelligence for health: WHO guidance.* World Health Organization, 2021.

10 *Artificial intelligence in healthcare: Applications, risks, and ethical and societal impacts.* European Parliamentary Research Service, 2022.

Table 1 | Glossary of key terms

Term	Definition
Artificial intelligence (AI)	The study and development of computer systems that can copy intelligent human behaviour.
Generative AI	The employment of AI to generate new content, such as text, audio, images, and video.
Large language model (LLM)	Generative AI that allows computers to understand and generate language content, similar to that of a human (eg, GPT-4, Gemini Pro).
Text-to-image model	Generative AI that allows computers to understand input text to generate image content.
Text-to-video model	Generative AI that allows computers to understand input text to generate video content.
Multimodal AI	AI systems capable of understanding and integrating diverse data types—text, audio, images, and video—to perform complex tasks with enhanced and diverse creative generation capabilities.
Generative AI application	Software platforms that incorporate generative AI to provide advanced capabilities for content creation and problem solving (eg, ChatGPT, Copilot, Midjourney, Sora).
AI safety	The broad discipline of ensuring AI systems and applications operate safely, avoiding misuse and harmful consequences. It focuses on developing norms, policies, governance, and technical measures to keep AI reliable and within desired parameters, aiming to mitigate risks to humans and the public.
Ethical AI	The discipline focused on ensuring AI development and use are fair, transparent, and accountable. Ethical AI mandates companies to adhere to ethical standards, prioritising societal welfare and individual rights.
AI vigilance	Analogous to pharmacovigilance (the formal systems for monitoring adverse effects of medicines), AI vigilance refers to implementing effective oversight, auditing, and transparency systems for AI to manage their evolving risks and impacts.
AI guardrails	Technical constraints implemented in AI applications that aim to maintain ethical and safe operations. They involve fine tuning models and applying front or back ended filters to prevent unintended or harmful outputs.
AI hallucinations	Instances where AI systems generate false or misleading information presented as if it was true, often with an apparent high level of confidence. A frequent cause is insufficient or biased training data.
Disinformation	The intentional creation and dissemination of false or misleading information, typically carried out by malicious actors to deceive, manipulate, or cause harm.