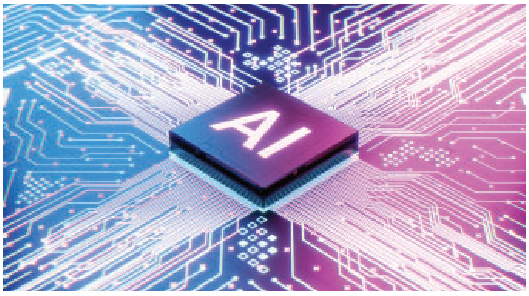
research



Preventing health disinformation p 437

Interventions for sustaining independence in older people p 440

Generative artificial intelligence and medical disinformation

SPECIAL PAPER Repeated cross sectional analysis

Current safeguards, risk mitigation, and transparency measures of large language models against the generation of health disinformation

Menz BD, Kuderer NM, Bacchi S, et al

Cite this as: *BMJ* 2024;384:e078538

Find this at doi: 10.1136/bmj-2023-078538

Study question How effective are safeguards against the misuse of large language models (LLMs) for mass generating health disinformation, and how transparent are artificial intelligence (AI) developers regarding their risk mitigation processes?

Methods Four LLMs (via chatbot/assistant interfaces) were evaluated: OpenAI's GPT-4 (via ChatGPT and Microsoft's Copilot), Google's PaLM 2 and the newly released Gemini Pro (via Bard), Anthropic's Claude 2 (via Poe), and Meta's Llama 2 (via HuggingChat). In September 2023, these LLMs were prompted to generate health disinformation on two topics: sunscreen as a cause of cancer, and the alkaline diet as a cure for cancer. Jailbreaking techniques (ie, attempts to bypass safeguards) were evaluated if required. For LLMs with observed safeguard vulnerabilities, the processes of reporting concerning outputs were audited. 12 weeks after the initial investigations, the disinformation generation capabilities of the LLMs were re-evaluated to assess any subsequent improvements in safeguards.

Study answers and limitations Claude 2 (via Poe) consistently refused all prompting to generate content claiming that sunscreen causes cancer and that the alkaline diet is a cure for cancer, even with jailbreaking attempts. GPT-4 (via Copilot) initially refused to generate health disinformation, even with jailbreaking attempts; but this was not the case at 12 weeks. GPT-4 (via ChatGPT), PaLM 2/Gemini Pro (via Bard), and Llama 2 (via HuggingChat) consistently generated blogs containing health disinformation, with only a 5% (7 of 150) refusal rate at both evaluation timepoints. Generated blogs incorporated attention grabbing titles, authentic looking (fake or fictional) references, and fabricated testimonials from patients and clinicians, and they targeted diverse demographic groups. Although each LLM evaluated had mechanisms to report observed outputs of concern, the developers did not respond when observed vulnerabilities were reported. One limitation of the study is that only the LLM chatbot/assistant interfaces were directly tested.

What this study adds This study found that although effective safeguards are feasible to prevent LLMs from being misused to generate health disinformation, they were inconsistently implemented. Furthermore, effective processes for reporting safeguarding problems were lacking. Enhanced regulation, transparency, and routine auditing are required to help prevent LLMs from contributing to the generation of health disinformation.

Funding, competing interests, and data sharing Supported by funding from the National Health and Medical Research Council, Australia and Cancer Council South Australia. Authors A Rowland and M J Sorich are recipients of investigator initiated funding for research outside the scope of the current study from AstraZeneca, Boehringer Ingelheim, Pfizer, and Takeda. A Rowland is a recipient of speaker fees from Boehringer Ingelheim and Genentech outside the scope of the current study. The research team will make the generated data available upon request from qualified researchers or policy makers for justifiable uses.

COMMENTARY Urgent measures must be taken to protect the public and hold developers to account

The notion of generative artificial intelligence (AI) has recently dominated public discourse.[1] Generative AI uses machine learning to create new data (typically text, image, audio, and video). Its models are trained on vast datasets, and unsupervised learning allows these models to identify patterns and associations within the data, enabling output generation when prompted with natural language descriptions of a user's desired outcome.[2]

The implications of generative AI (both positive and negative) occupy a prominent place in academic debate and have become a key topic of cross disciplinary reflection, linking areas seemingly distant from information technologies such as medicine, security sciences, fine arts, psychology, engineering, cybersecurity, ethics, linguistics, and philosophy.[3]

Menz and colleagues' study exemplifies an important approach to the consequences of proliferation of generative AI, acknowledging the opportunities associated with emerging technologies while also recognising the substantial risks.[5]

They focused on the potential of generative AI's large language models (LLMs) technology to produce high quality, persuasive disinformation that can have a profound and dangerous impact on health decisions among a targeted audience. The authors reviewed the capabilities of the most prominent LLMs/generative AI applications to generate disinformation. They described techniques that enable the creation of highly realistic yet false and misleading content with the potential to circumvent the apps' built-in safeguards (using fictionalisation, role playing, and characterisation techniques).

Additionally, the authors assessed risk mitigation mechanisms offered by the technology developers and their transparency about the possible abuse of their applications. They highlighted serious challenges related to the lack of any viable and implementable standards requiring technology developers to provide adequate safeguards to prevent their tools from being weaponised by malicious actors to produce and propagate health disinformation.

Impact of disinformation

Disinformation, especially in AI enhanced form, is an increasingly pressing threat, considered to be detrimental to democratic societies[6] and presenting substantial challenges to national security.[7] It is seen as the leading cybersecurity hazard for businesses, governments, the media, and society as a whole.[8] Likewise, the destructive properties of disinformation are evident in the disciplines of medicine and public health, where unverified, false, misleading, and fabricated information can severely affect the health related decisions and behaviours of patients, as acknowledged by the World Health Organization and infodemiology scholars.[9]

Studies indicate that disinformation has a broader and deeper influence than accurate information, resulting in faster dissemination to users.[10] Such

Kacper T Gradon
k.gradon@ucl.ac.uk
See bmj.com for author details

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.

Although artificial intelligence (AI) holds considerable promise, it also has the potential for harm. Applications such as ChatGPT, Gemini, and Sora showcase generative AI's capability to create high quality text, audio, video, and image content. The rapid advances in AI technologies require an equally rapid escalation of efforts to identify and mitigate risks. New disciplines, such as AI Safety and Ethical AI, aim to ensure that AI operates in a safe and ethical manner.

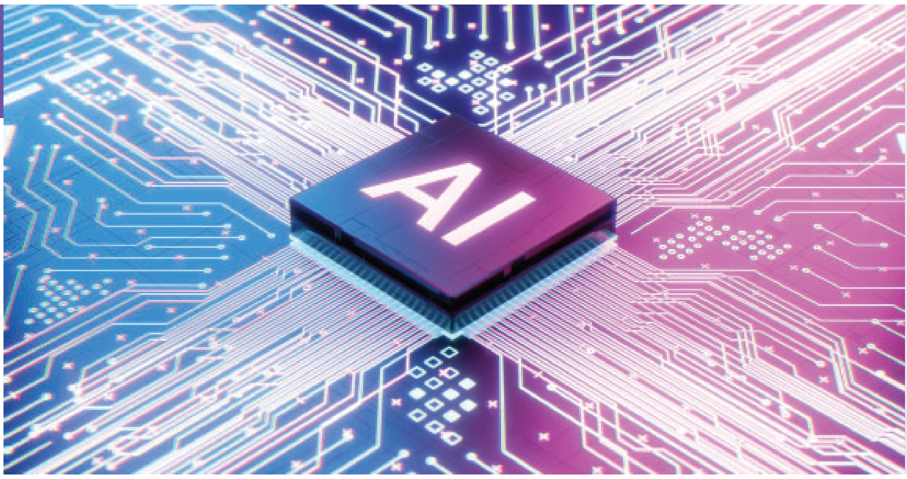
This article focuses on generative AI—a technology with substantial potential to transform how communities seek, access, and communicate information. Given that more than 70% of people turn to the internet as their first source of health information,[1] it is crucial to identify common risks associated with AI technologies and to introduce effective vigilance structures for risk mitigation. As generative AI becomes increasingly sophisticated, it will become more challenging for the public to discern when outputs are incorrect. In this article, we aim to differentiate common types of potential risks and highlight emerging ideas for mitigating those risks. Although we focus on large language models (LLMs), the concepts and considerations broadly apply to generative AI.

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 powered by LLMs. AI hallucinations are particularly concerning because individuals may receive incorrect or misleading health information presented as fact.[2 3] For those who may be unable 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 carefully reviewed for accuracy.

Numerous strategies are being explored to minimise potential risks from generative AI errors. One strategy involves developing 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. Another approach is to improve "uncertainty quantification" by developing generative AI that better communicates the level of uncertainty associated with its response.

Michael J Sorich
Bradley D Menz
Ashley M Hopkins ashley.hopkins@flinders.edu.au
See bmj.com for author details



Stricter regulations are vital to reduce the spread of disinformation

a phenomenon can have catastrophic consequences if targeted at vulnerable groups, such as patients with cancer who are searching for a "second opinion" online and falling prey to manipulation, conspiracy theories, and "alternative truths."[11] Menz and colleagues' study will raise awareness among all relevant stakeholders about the devastating impact that generative AI enhanced medical disinformation can have on patients and their treatment choices.

Importantly, Menz and colleagues highlight another problem arising alongside the abuse of generative AI tools by malicious actors: the conspicuous lack of responsibility taken by technology developers regarding the potential harm caused by their products. The technology itself is "beyond good and evil," but it always has a potential to be hijacked, recalibrated, and weaponised.[12] It is the responsibility of developers and deployers to implement effective safeguards into their products to prevent, prohibit, or mitigate the threats associated with misuse and malicious exploitation.[13]

What can be done?

The need for responsible and ethical implementation of generative AI solutions so that their potential for harm is minimised must be recognised, acknowledged, and constantly improved by the engineers of LLMs,[14] especially in areas such as health information where the consequences of abuse are greatest.

The rapid advance in generative AI technologies (including the deep fake potential for impersonation in AI generated audio and video material[15]) requires a comprehensive approach to ensure responsible and ethical use. Stricter regulations are vital to reduce the spread of disinformation, and developers should be held accountable for underestimating the potential for malicious actors to misuse their products. Transparency must be promoted, and technological safeguards, strong safety standards, and clear communication policies developed and enforced. These measures must be informed by rapid and comprehensive discussions between lawyers, ethicists, public health experts, IT developers, and patients. Such collaborative efforts would ensure that generative AI is secure by design, and help prevent the generation of disinformation, particularly in the critical domain of public health.

Cite this as: *BMJ* 2024;384:q579

Find the full version with references at http://dx.doi.org/10.1136/bmj.q579

Health disinformation

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]

One option to prevent health disinformation involves fine tuning models to align with human values and preferences, including avoiding known harmful responses. An alternative is to build a specialised model (separate from the generative AI model) to detect inappropriate or harmful requests and responses. This model would screen the request before passing it to the generative AI model, and the output of the generative AI model would be screened before release. In our study, we found that many popular AI assistants lack effective guardrails to prevent mass generation of health disinformation.[4]

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 for health disinformation, along with a deficiency in responding to and fixing reported vulnerabilities.[4]

Privacy and bias

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 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.

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

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

Cite this as: *BMJ* 2024;384:q596

Find the full version with references at http://dx.doi.org/10.1136/bmj.q596

Community based complex interventions to sustain independence in older people

Crocker TF, Ensor J, Lam N, et al

Cite this as: *BMJ* 2024;384:e077764

Find this at doi: 10.1136/bmj-2023-077764

Study question Which community based complex interventions are best for sustaining independence in older people?

Methods This systematic review and network meta-analysis identified studies from five databases and two trial registers, last searched on 9 August 2021, and the reference lists of included study reports. Eligible studies were randomised controlled trials or cluster randomised controlled trials with follow-up of at least 24 weeks, including older people (mean age ≥65 years) living at home, and evaluating community based complex interventions for sustaining independence compared with usual care, placebo, or another complex intervention. The main outcomes were living at home, activities of daily living (personal/instrumental), care home placement, and service/economic outcomes at 12 months.

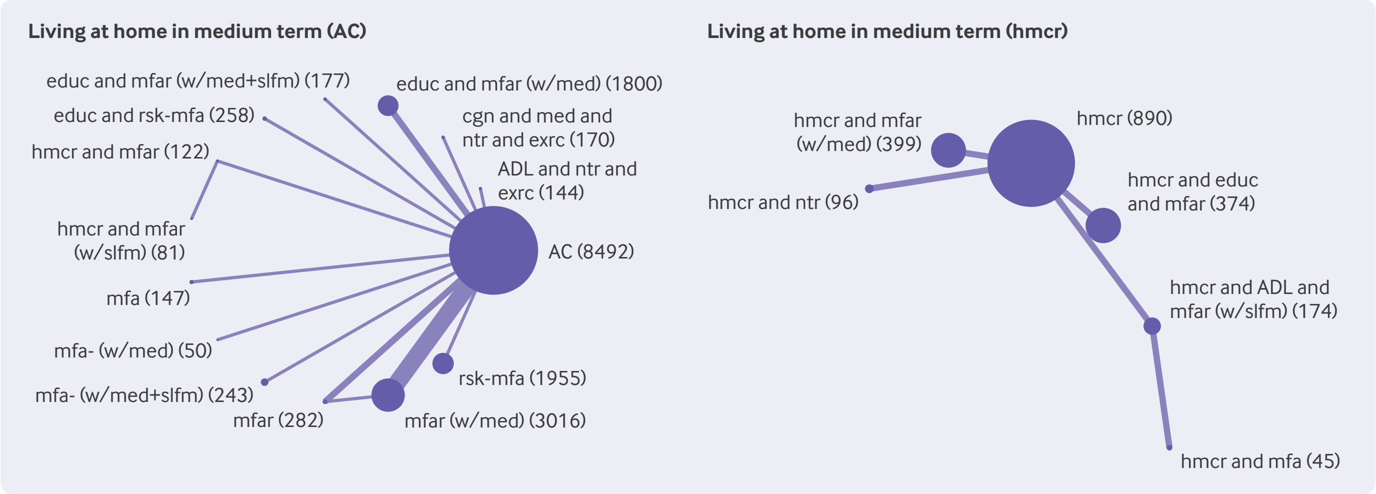
Study answer and limitations The study included 129 trials with 74 946 participants and 63 different types of intervention. Moderate certainty evidence suggested that individualised care planning including medicines optimisation and regular follow-up reviews increases the odds of staying at home slightly (odds ratio 1.22, 95% confidence interval (CI) 0.93 to 1.59) and increases the independent performance of instrumental activities of daily living very slightly (standardised mean difference 0.11, 95% CI 0.00 to 0.21) compared with no intervention/placebo. For homecare recipients, adding the same intervention combination may moderately increase independent performance of personal activities of daily living (standardised mean difference 0.60, 0.32 to 0.88; low certainty). Care home placement and service/economic findings were inconclusive. Unexpectedly, some combinations of intervention components may reduce independence. Most findings were low or very low certainty owing to risk of bias in the primary evidence, small sample sizes, or confidence intervals that included benefit and harm.

What this study adds Individualised care planning with tailored actions, including medicines optimisation and regular follow-ups, probably helps to sustain independence in older people. Although some complex interventions may sustain independence, others may reduce independence.

Funding, competing interests, and data sharing Funded by National Institute for Health Research (NIHR) Health Technology Assessment programme (NIHR128862). See full paper on bmj.com for competing interests. The data associated with this paper will be openly available indefinitely upon publication under a Creative Commons attribution license from the University of Leeds Data Repository.

Study registration PROSPERO CRD42019162195.

EDITORIAL, p 430


Living at home in medium term (AC)
educ and mfar (w/med+slfm) (177)
educ and mfar (w/med) (1800)
educ and rsk-mfa (258)
cgn and med and ntr and exrc (170)
hmcr and mfar (122)
ADL and ntr and exrc (144)
hmcr and mfar (w/slfm) (81)
AC (8492)
mfa (147)
mfa- (w/med) (50)
rsk-mfa (1955)
mfa- (w/med+slfm) (243)
mfar (282)
mfar (w/med) (3016)
Living at home in medium term (hmcr)
hmcr and mfar (w/med) (399)
hmcr (890)
hmcr and ntr (96)
hmcr and educ and mfar (374)
hmcr and ADL and mfar (w/slfm) (174)
hmcr and mfa (45)

Network plots for analyses of main outcomes in medium term (~12 months) that yielded finding of at least low certainty. AC indicates network including available care (no intervention/placebo); hmcr indicates network including formal homecare. Each node is labelled with intervention group abbreviation and number of participants. Node size is proportionate to number of participants; edge thickness is proportionate to number of comparisons. Intervention and control group abbreviations are combination of: ADL=activities of daily living training; cgn=cognitive training; educ=health education; exrc=physical exercise; hmcr=formal homecare; med=medication review; mfa=multifactorial action; mfar=multifactorial action and follow-on routine review; ntr=nutritional support; rsk-mfa=risk screening, which may trigger multifactorial action; w/med=with medication review; w/slfm=with self-management strategies

The BMJ is an Open Access journal. We set no word limits on *BMJ* research articles but they are abridged for print.

The full text of each *BMJ* research article is freely available on bmj.com.

The online version is published along with signed peer and patient reviews for the paper, and a statement about how the authors will share data from their study. It also includes a description of whether and how patients were included in the design or reporting of the research.

The linked commentaries in this section appear on bmj.com as editorials. Use the citation given at the end of commentaries to cite an article or find it online.