Department of Cybersecurity, Faculty of Electronics and Information Technology, Warsaw University of Technology, Warsaw, Poland

k.gradon@ucl.ac.uk

Cite this as: *BMJ* 2024;384:q579
http://dx.doi.org/10.1136/bmj.q579

Generative artificial intelligence and medical disinformation

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

Kacper T Gradon *associate professor*

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] It is difficult to find a specialty that ignores the potential impact of generative AI on the functioning of individuals and social groups, or humanity in general.[4]

The linked paper by Menz and colleagues (doi:10.1136/bmj-2023-078538) 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]

In their study, Menz and colleagues 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.

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 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]

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 indifference, lack of transparency, and unresponsiveness of generative AI companies to deal with the vulnerabilities of their own inventions is one of the more disturbing aspects of Menz and colleagues' study.

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.

Competing interests: The BMJ has judged that there are no disqualifying financial ties to commercial companies. The author does not declare any other interests.

Further details of The BMJ policy on financial interests is here: https://www.bmj.com/sites/default/files/attachments/resources/2016/03/16-current-bmj-education-coi-form.pdf

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

1 Miyazaki K, Murayama T, Uchiba T, An J, Kwak H. Public perception of generative AI on Twitter: an empirical study based on occupation and usage. *EPJ Data Sci* 2024;13:. doi: 10.1140/epjds/s13688-023-00445-y.

2 Oniani D, Hilsman J, Peng Y, etal. Adopting and expanding ethical principles for generative artificial intelligence from military to healthcare. *NPJ Digit Med* 2023;6:. doi: 10.1038/s41746-023-00965-x. pmid: 38042910

3 Baldassarre MT, Caivano D, Fernandez Nieto B, Gigante D, Ragone A. The Social Impact of Generative AI: An Analysis on ChatGPT. Proceedings of the 2023 ACM Conference on Information Technology for Social Good; 2023; GoodIT'23, Sept 23:363-73. doi: 10.1145/3582515.3609555

4 Sabherwal R, Grover V. The Societal Impacts of Generative Artificial Intelligence: A Balanced Perspective. *J Assoc Inf Syst* 2024;25:-22. doi: 10.17705/1jais.00860.

5 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.

6 Kreps S, Kriner D, How AI. Threatens Democracy. *J Democracy* 2023;34:-31. doi: 10.1353/jod.2023.a907693.

7 Sarts J. Disinformation as a Threat to National Security. In: Jayakumar S, Ang B, Anwar ND, eds. *Disinformation and Fake News*. Palgrave Macmillan, 2021: -35, doi: 10.1007/978-981-15-5876-4_2.

8 Mazurczyk W, Lee D, Vlachos A. Disinformation 2.0 in the Age of AI: A Cybersecurity Perspective. *Commun ACM* 2024;67:-9. doi: 10.1145/3624721.

9 Calleja N, AbdAllah A, Abad N, etal. A Public Health Research Agenda for Managing Infodemics: Methods and Results of the First WHO Infodemiology Conference. *JMIR Infodemiology* 2021;1:e30979. doi: 10.2196/30979

10 Gradon KT, Holyst JA, Moy WR, Sienkiewicz J, Suchecki K. Countering misinformation: A multidisciplinary approach. *Big Data Soc* 2021;8. doi: 10.1177/20539517211013848.

11 Swire-Thompson B, Johnson S. Cancer: A model topic for misinformation researchers. *Curr Opin Psychol* 2023;56:101775. doi: 10.1016/j.copsyc.2023.101775. pmid: 38101247

12 Gradon KT. Electric sheep on the pastures of disinformation and targeted phishing campaigns. The security implications of ChatGPT. *IEEE Secur Priv* 2023;21. doi: 10.1109/MSEC.2023.3255039.

13 World Health Organization. Ethics and governance of artificial intelligence for health. Guidance on large multi-modal models. Geneva: World Health Organization; 2024. https://www.who.int/publications/i/item/9789240084759 (accessed 4 March 2024).

14 Widder DG, Nafus D. Dislocated accountabilities in the "AI supply chain": Modularity and developers' notions of responsibility. *Big Data Soc* 2023;10. doi: 10.1177/20539517231177620.

15 Patil S. Artificial Intelligence face swapping: promise and peril in healthcare. *Mayo Clin Proc Digit Health* 2024;2:. doi: 10.1016/j.mcpdig.2024.01.009.