BJPsych | The British Journal of Psychiatry (2024)
224, 33–35. doi: 10.1192/bjp.2023.136

BJPsych Editorial

Artificial intelligence and increasing misinformation

Scott Monteith, Tasha Glenn, John R. Geddes, Peter C. Whybrow, Eric Achtyes and Michael Bauer

Summary

With the recent advances in artificial intelligence (AI), patients are increasingly exposed to misleading medical information. Generative AI models, including large language models such as ChatGPT, create and modify text, images, audio and video information based on training data. Commercial use of generative AI is expanding rapidly and the public will routinely receive messages created by generative AI. However, generative AI models may be unreliable, routinely make errors and widely spread misinformation. Misinformation created by generative AI about mental illness may include factual errors, nonsense, fabricated sources and dangerous advice. Psychiatrists need to recognise that patients may receive misinformation online, including about medicine and psychiatry.

Keywords

Generative artificial intelligence; misinformation; artificial intelligence; disinformation; technology.

Scott Monteith is a psychiatrist, Psychiatry Clerkship Director at Michigan State University and Associate Program Director of Pine Rest's Rural Track Psychiatry Residency, Michigan, USA. **Tasha Glenn** is Director of the non-profit ChronoRecord Association, California, USA. **John R. Geddes** is a WA Handley Professor of Psychiatry and Director of the National Institute for Health and Care Research (NIHR) Oxford Health Biomedical Research Centre, UK. **Peter C. Whybrow** is Professor of Psychiatry at the Semel Institute for Neuroscience and Human Behavior, University of California Los Angeles (UCLA), USA. **Eric Achtyes** is a professor and Chair of the Department of Psychiatry at the Western Michigan University Homer Stryker M.D. School of Medicine, USA. **Michael Bauer** is Professor of Psychiatry and Chair of the Department of Psychiatry and Director of the Psychiatric University Hospital, Technische Universität Dresden, Germany.

Although there is widespread excitement about the creative successes and new opportunities resulting from the recent transformative technological advancements in artificial intelligence (AI), one result is increasing patient exposure to medical misinformation. We now live in an era of synthetic media. Text, images, audio and video information can be created or altered by generative AI models based on the data used to train the model. The commercial use of automated content produced by generative AI models, including large language models (LLMs) such as ChatGPT, GPT-3 and image generation models, is expanding rapidly. Private industry, not academia, is dominating the development of the new AI technology.[1] The potential business applications for generative AI models are wide-ranging: creating marketing and sales copy, product guides and social media posts, sales support chatbots for customers, software development and human resources support. But generative AI models such as ChatGPT can be unreliable, making errors of both fact and reasoning that can be spread on an unprecedented scale.[2] The general public can easily get incorrect information from generative AI on any topic, including medicine and psychiatry. The spread of misinformation created by generative AI can be accelerated by unsuspecting acceptance of content accuracy. There are serious potential negative consequences of medical misinformation relating to individual care as well as public health. Psychiatrists need to be aware of the rapid spread of misinformation online.

Introduction to generative AI

The focus of traditional AI is on predictive models to perform a specific task, such as estimate a number, classify data or select between a set of options. In contrast, the focus of generative AI is to create original content. For a given input, rather than one correct answer based on the model's decision boundaries, generative AI models produce text, audio and visual outputs that can easily be mistakenly attributed to human authors.

Generative AI models are based on large neural networks that are trained using an immense amount of raw data.[3] Three major factors have contributed to the recent advancements in generative models: the explosion of training data now available on the internet, improvements in training algorithms and increases in available computing power for training the models.[3] For example, GPT-3 was trained using an estimated 45 terabytes of text data, or about 1 million feet of bookshelf space.[4] The training process broke the text into pieces of words called tokens and created 175 million parameters that generate new text by statistically identifying the most probable next token in a sequence of tokens.[5] A newer version of GPT-4 is a multimodal LLM, responding to both text and video images.

Generative AI can create the illusion of intelligence. Although at times the output of generative AI models can seem astonishingly human-like, they do not understand the meaning of words and frequently make errors of reasoning and fact.[2,5] The statistical patterns determine the word sequences without any understanding of the meaning or context in the real world.[5] Researchers in the generative AI field often use the word 'hallucination' to describe output generated by LLM that is nonsensical, not factual, unfaithful to the underlying content, misleading, or partially or totally incorrect. The many types of error from generative AI models include factual errors, inappropriate or dangerous advice, nonsense, fabricated sources and arithmetical errors. Other issues include outdated responses reflecting the year that LLM training occurred, and different answers to iterations of the same question. One example of inappropriate or dangerous advice is a chatbot recommending calorie restriction and dieting after being told the user has an eating disorder.[6]

https://doi.org/10.1192/bjp.2023.136 Published online by Cambridge University Press

The output of generative AI models may contain toxic language, including hate speech, insults, profanity and threats, despite some efforts at filtering. The fundamental problem is the prevalence of biases in the internet data used for training generative AI models related to race/ethnicity, gender and disability status. Although human feedback is being used to score responses and improve the safety of generative AI models, biases remain. Another concern is that the output of generative AI models may contain manipulative language since internet data also contain a vast amount of manipulative content.

Attitudes to generative AI

In addition to widespread commercial expansion, generative AI, and ChatGPT in particular, is extremely popular with the general public. AI products, including generative AI, are routinely anthropomorphised, or described and characterised as having human traits, by the general public, media and AI researchers. It is easy for the general public to anthropomorphise the use of LLMs, given the simplicity of conversing and the authoritative-sounding responses. The media routinely describe LLMs using words suggestive of human intelligence, such as ‘thinks’, ‘believes’ and ‘understands’. These portrayals generate public interest and trust, but also downplay the limitations of LLMs that statistically predict word sequences based on patterns learned from the training data. Researchers also anthropomorphise generative AI, referring to undesirable LLM text errors as ‘hallucinations’. Since the general public will associate hallucinations with unreal human sensory perceptions, this word may imply a false equivalency between LLMs and the human mind.

Incorrect output from generative AI models often seems plausible to many people, especially those unfamiliar with the topic. A major problem with generative AI is that people who do not know the correct answer to a question will not be able to tell if an answer is wrong.[7] Human intelligence is needed to evaluate the accuracy of generative AI output.[7] Although generative AI products are improving, so is the ability to create outputs that sound convincing but are incorrect.[7] Many people do not realise how often generative AI models are incorrect. People are unaware that unless they are experts in the field, they must carefully check the answers to questions, even if the text sounds very convincing.

Intentional spread of misinformation

Generative AI models enable the automation and rapid dissemination of intentional misinformation campaigns.[3] LLM products can automate the intentional creation and spread of misinformation on an extraordinary scale.[2,3] Without having to rely on human labour, the automated generation of misinformation drives down the cost of creating and disseminating misinformation. Misinformation created by the generative AI models may be better written and more compelling than that from human propagandists. The spread of online misinformation in all areas of medicine is particularly dangerous.

In addition to knowledge of the subject area, an individual’s understanding of technology and online habits will affect their acceptance and spreading of misinformation. People may be in the habit of sharing news on social media or be overly accepting of online claims. Some people with mental illness may be especially vulnerable to online misinformation. Generative AI products will further increase the volume of information shared, including on medical topics. The use of generative AI emphasises the need and importance of increasing digital training opportunities for the general public from validated sources.

Unique ethical issues

In addition to accuracy, reliability, bias and toxicity, there are many unsettled ethical and legal issues related to generative AI. There are privacy issues related to the collection and use of personal and proprietary data for training models without permission and compensation. There are legal issues that include plagiarism, copyright infringement and responsibility for errors and false accusations in generative AI output.

Conclusions

The use of generative AI products in commerce, healthcare and by the general public is rapidly growing. In addition to beneficial uses, there are serious potential negative impacts from AI-generated and widely spread misinformation. The misinformation created by generative AI about mental illness may include factual errors, nonsense, fabricated sources and dangerous advice. Measures to mitigate the dangers of misinformation from generative AI need to be explored. Psychiatrists should realise that patients may be obtaining misinformation and making decisions based on generative AI responses in medicine, and many other topics, that may affect their lives.

Scott Monteith, Michigan State University College of Human Medicine, Traverse City Campus, Traverse City, Michigan, USA; **Tasha Glenn**, ChronoRecord Association, Fullerton, California, USA; **John R. Geddes**, Department of Psychiatry, University of Oxford, Warneford Hospital, Oxford, UK; **Peter C. Whybrow**, Department of Psychiatry and Biobehavioral Sciences, Semel Institute for Neuroscience and Human Behavior, University of California Los Angeles (UCLA), Los Angeles, California, USA; **Eric Achtyes**, Department of Psychiatry, Western Michigan University Homer Stryker M.D. School of Medicine, Kalamazoo, Michigan, USA; **Michael Bauer**, Department of Psychiatry and Psychotherapy, University Hospital Carl Gustav Carus Medical Faculty, Technische Universität Dresden, Dresden, Germany

Correspondence: Scott Monteith. Email: monteit2@msu.edu

First received 15 Jun 2023, final revision 6 Sep 2023, accepted 20 Sep 2023

Data availability

Data availability is not applicable to this article as no new data were created or analysed in this study.

Author contributions

S.M. and T.G. wrote the initial draft. All authors reviewed and approved the final manuscript.

Funding

This work received no specific grant from any funding agency, commercial or not-for-profit sectors

Declaration of interest

J.R.G., Director of the NIHR Oxford Health Biomedical Research Centre, is a member of the *BJPsych* editorial board and did not take part in the review or decision-making process of this paper.

https://doi.org/10.1192/bjp.2023.136 Published online by Cambridge University Press

References

1 Ahmed N, Wahed M, Thompson NC. The growing influence of industry in AI research. *Science* 2023; **379**: 884–6.

2 Marcus G. AI platforms like ChatGPT are easy to use but also potentially dangerous. *Sci Am* 2022; **31**: 19 Dec (https://www.scientificamerican.com/article/ai-platforms-like-chatgpt-are-easy-to-use-but-also-potentially-dangerous/).

3 Goldstein JA, Sastry G, Musser M, DiResta R, Gentzel M, Sedova K. Generative language models and automated influence operations: emerging threats and potential mitigations. *arXiv* [preprint] 2023. Available from: https://doi.org/10.48550/arXiv.2301.04246.

4 McKinsey & Co. *What is Generative AI?* McKinsey & Co, 2023 (https://www.mckinsey.com/~/media/mckinsey/featured%20insights/mckinsey%20explainers/what%20is%20generative%20ai/what%20is%20generative%20ai.pdf).

5 Smith GN. *Large Learning Models Are an Unfortunate Detour in AI*. Mind Matters, 2022 (https://mindmatters.ai/2022/12/large-learning-models-are-an-unfortunate-detour-in-ai/).

6 Bailey C. Eating disorder group pulls chatbot sharing diet advice. *BBC News* 2023: 1 Jun (https://www.bbc.com/news/world-us-canada-65771872).

7 Narayanan A, Kapoor S. ChatGPT is a bullshit generator. But it can still be amazingly useful. *AI Snake Oil* 2022: 6 Dec (https://aisnakeoil.substack.com/p/chatgpt-is-a-bullshit-generator-but).

Psychiatry in literature

Personality and character

George Ikkos

In Classical Greek πρόσωπον means face or mask. The early Latin equivalent is *persona* and contemporary derivatives include personality, even parson. Personality therefore alludes to the face we show the world – what we bare, veil and exaggerate. Dictionaries often conflate personality and character but we may discern differences. While celebrities can be ‘personalities’, actors portray ‘characters’. The word ‘character’ has its roots in the Greek word χαρακτήρας, meaning ‘engraved mark’ or ‘instrument for marking’. A cutting through of a kind. Confronted with acute dilemmas we may act ‘out of character’ so to say, thus show character and make our mark!

The British Journal of Psychiatry (2024) 224, 35. doi: 10.1192/bjp.2023.172

https://doi.org/10.1192/bjp.2023.136 Published online by Cambridge University Press