GENERATIVE AI

One in a series of discussion primers about tricky topics in human research ethics.

An evolving space

The term "Generative Artificial Intelligence" refers to a family of digital tools that are trained on very large data sets and use machine learning techniques to produce text, images, code or other media in response to an input or stimulus.

There are now thousands of examples of Generative AI tools, including ChatGPT, Bard, Bing, LLaMA, Stable Diffusion, MidJourney, DALL-E, and GitHub CoPilot.

Recent advances in Generative AI carry both opportunities and risks for research. This primer is designed to promote discussion between ethics committee members and other reviewers.

Have you used any Generative AI tools? If so, which tools have you used, and for what purposes? What's your general impression of these tools?

Scary opportunities

Generative AI tools present opportunities for impactful research and innovation.

Consider an April 2023 study by Ayers et al. which aimed to understand whether an AI chatbot could provide good medical advice. The research team randomly chose 195 patient questions from a social media forum where a verified physician had responded to the question. The research team then asked an AI chatbot the same questions. The anonymised answers were evaluated in triplicate by a team of licensed health care professionals.

It turned out that the expert evaluators preferred the chatbot responses in 78.6% of cases. They judged that the AI responses were not only of higher factual quality, but also more empathetic. [1]

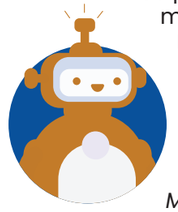
Imagine that researchers now propose to trial this chatbot (discussed above) with patients who have serious medical conditions, and evaluate their health outcomes over a period of six months. Half the participants will receive medical advice from a human physician; the other half will receive advice exclusively from a chatbot. What do you think are the ethical issues here? Does the National Statement provide any helpful guidance?

Information and consent

Some AI tools help researchers complete existing research processes more efficiently — for example, automatically transcribing audio or video, or grouping text-based survey responses into themes. To what extent do you think researchers should inform participants about their use of these tools?

Put yourself in participants' shoes. How would you feel if researchers were uploading audio recordings of your voice to an AI transcription service? Would you personally have any concerns? More broadly, do you think there are research studies for which some participants might not consent precisely *because* researchers are using emerging AI tools?

Justice, bias, inequality



One known issue with Generative AI tools is that they can be prone to systematic bias. This may be due to the models that they employ and/or the data on which they are trained.

For example, Thomas and Thomson investigated how MidJourney returned AI-generated images of journalists in response to different inputs. They found that, "*[f]or non-specialised job titles, Midjourney returned images of only younger men and women. For specialised roles, both younger and older people were shown – but the older people were always men.*" [2] Similarly, Luccione, el al. found that DALL-E, "*generated white men 97% of the time when given prompts like "CEO" or "director."* [3]

Tricky Goose Training asked MidJourney for a "Simple vector image of a CEO of a large company", and these were the four options returned on the first request.

A 2019 study found evidence of bias in an AI tool used to identify patients for follow-up care. *"The authors estimated that this racial bias reduces the number of Black patients identified for extra care by more than half."* [4]

A related issue is that the large companies that own popular generative AI tools tend to be secretive about their models and training data, making it nearly impossible for researchers to assess the extent to which the outputs may be skewed.

On the other hand, some scholars have argued that AI tools could potentially help *avoid* human bias in some cases — for example in coding for qualitative studies. "*By automating the coding and categorization process, ChatGPT diminishes the risk of coder bias, thereby amplifying the reliability of research results.*" [5]

After consulting the papers referenced above, what do you think? Could the use of generative AI tools translate to issues with merit or justice for research studies that you review? What could researchers do to address and mitigate those risks?

Privacy and security

The National Statement requires that "*researchers and their institutions should respect the privacy, confidentiality and cultural sensitivities of the participants and, where relevant, of their communities.*" [NS 1.11] How can participants' privacy be protected while using generative AI tools? Many generative AI tools are cloud-based, and for some tools there are legitimate questions concerning their security. A confirmed breach of ChatGPT, for example, "led to the unexpected exposure of users' conversations..." [6]

How do you think researchers could preserve privacy while using generative AI? Have they considered locally hosted (rather than cloud-based) AI models? Would privacy concerns be sufficiently reduced if researchers removed identifiers from human data before uploading it to a generative AI tool? What evidence could researchers provide to demonstrate sufficient privacy protection? More generally, are privacy concerns in relation to generative AI tools any greater than they are for other online tools that researchers currently use? You may like to read and discuss reference [7].

Ethics applications drafted by AI?

Large Language Models such as ChatGPT are useful in drafting text in response to questions. Some academics are therefore hopeful that they can use these tools to reduce the burden of research administration. What about ethics applications? Could AI tools prompt researchers to consider ethical issues and risk mitigation strategies, write clear participant information documents, and draft responses to application questions in lay language? Could they also make the ethics application process more efficient? Or would any such use by researchers be just as unethical as the reported use by some who used AI to peer review national grant proposals? [8]

As a reviewer, how do you feel about the potential use of AI by researchers in writing their ethics applications? Would it be ok, as long as it was transparently disclosed?

References

[1] John W. Ayers, et al. **Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum**. JAMA Intern Med, 2023; DOI: 10.1001/jamainternmed.2023.1838

[2] Ryan J, Thomas and T. J. Thompson. **What does a Journalist Look Like? Visualizing Journalistic Roles through AI**. Digital Journalism, 2023; DOI: 10.1080/21670811.2023.2229883

[3] Alexandra Sasha Luccioni, et al. **Stable Bias: Analyzing Societal Representations in Diffusion Models**. arXiv, 2023; DOI: 10.48550/arXiv.2303.11408 [**Preprint; not peer reviewed**]

[4] Ziad Obermeyer, et al. **Dissecting racial bias in an algorithm used to manage the health of populations**. Science, 2019; DOI: 10.1126/science.aax2342

[5] Mohammed Salah, et al. **May the force of text data analysis be with you: Unleashing the power of generative AI for social psychology research**. Computers in Human Behavior: Artificial Humans, 2023; DOI: 10.1016/j.chbah.2023.100006

[6] Sue Poremba. **ChatGPT confirms data breach, raising security concerns**, Security Intelligence, 2023; URL: https://securityintelligence.com/articles/chatgpt-confirms-data-breach/

[7] Maanak Gupta, et al. **From ChatGPT to ThreatGPT: Impact of Generative AI in Cybersecurity and Privacy**, IEEE Access, 2023; DOI: 10.1109/ACCESS.2023.3300381

[8] Donna Lu. **Are Australian Research Council reports being written by ChatGPT?** The Guardian, 8 Jul 2023

TRICKY goose TRAINING