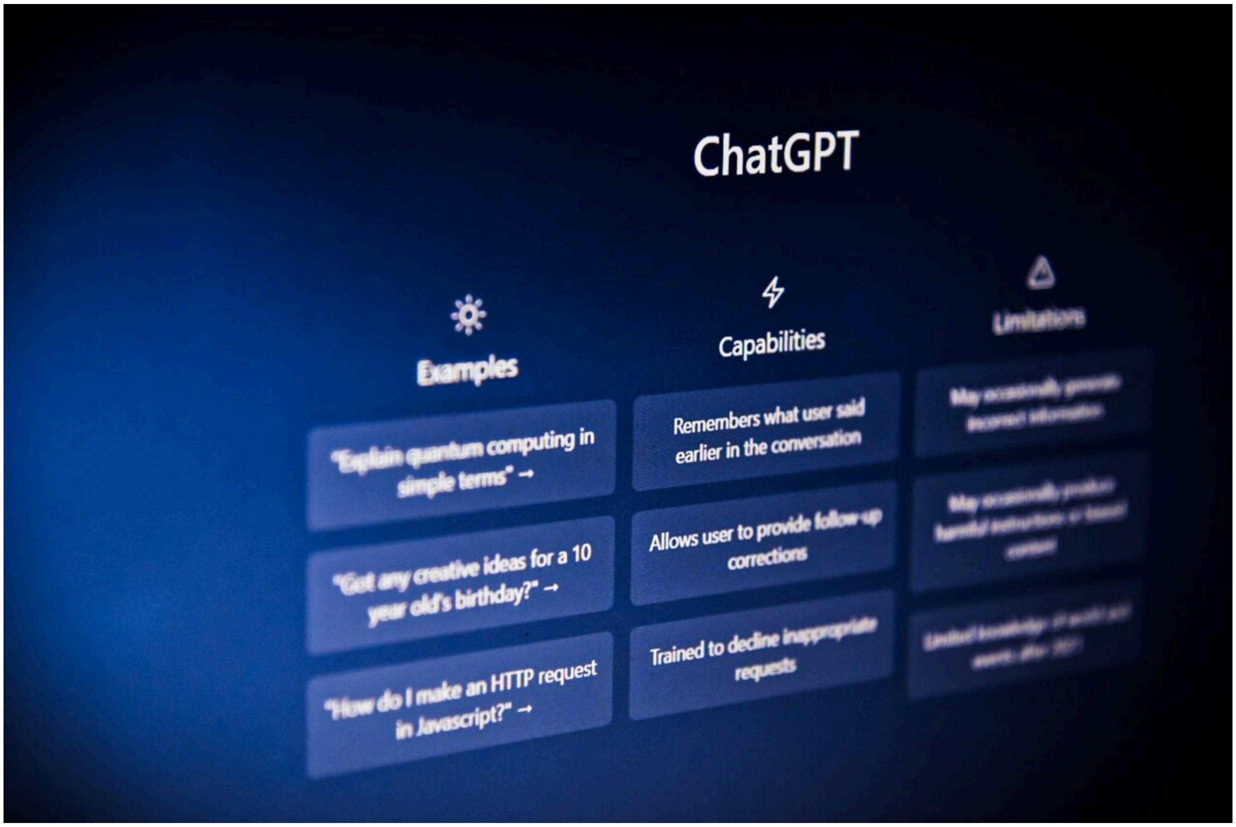
Generative AI models are encoding biases and negative stereotypes in their users, say researchers

June 22 2023


ChatGPT
Examples
"Explain quantum computing in simple terms" →
"Got any creative ideas for a 10 year old's birthday?" →
"How do I make an HTTP request in Javascript?" →
Capabilities
Remembers what user said earlier in the conversation
Allows user to provide follow-up corrections
Trained to decline inappropriate requests
Limitations
May occasionally generate incorrect information
May occasionally produce harmful instructions or biased content
Limited knowledge of world and events after 2021

Credit: Unsplash/CC0 Public Domain

In the space of a few months generative AI models, such as ChatGPT, Google's Bard and Midjourney, have been adopted by more and more people in a variety of professional and personal ways. But growing research is underlining that they are encoding biases and negative stereotypes in their users, as well as mass generating and spreading seemingly accurate but nonsensical information. Worryingly, marginalized groups are disproportionately affected by the fabrication of this nonsensical information.

In addition, mass fabrication has the potential to influence human belief as the models that drive it become increasingly common, populating the World Wide Web. Not only do people grab information from the web, but much of the primary training material used by AI models comes from here too. In other words, a continuous feedback loop evolves in which biases and nonsense become repeated and accepted again and again.

These findings—and a plea for psychologists and machine learning experts to work together very swiftly to assess the scale of the issue and devise solutions—are published today in a Perspective in the journal *Science*, which is co-authored by Abeba Birhane, who is an adjunct assistant professor in Trinity's School of Computer Science and Statistics (working with Trinity's Complex Software Lab) and Senior Fellow in Trustworthy AI at the Mozilla Foundation.

Prof Birhane said, "People regularly communicate uncertainty through phrases such as 'I think,' response delays, corrections, and speech disfluencies. By contrast, generative models give confident, fluent responses with no uncertainty representations nor the ability to communicate their absence. As a result, this can cause greater distortion compared with human inputs and lead to people accepting answers as factually accurate. These issues are exacerbated by financial and liability interests incentivizing companies to anthropomorphize generative models as intelligent, sentient, empathetic, or even childlike."

One such example provided in the Perspective focuses on how statistical regularities in a model assigned Black defendants with higher risk scores. Court judges, who learned the patterns, may then change their sentencing practices in order to match the predictions of the algorithms. This basic mechanism of statistical learning could lead a judge to believe Black individuals to be more likely to re-offend—even if use of the system is stopped by regulations like those recently adopted in California.

Of particular concern is the fact that it is not easy to shake biases or fabricated information once it has become accepted by an individual. Children are at especially high risk as they are more vulnerable to belief distortion as they are more likely to anthropomorphize technology and are more easily influenced.

What is needed is swift, detailed analysis that measures the impact of generative models on human beliefs and biases.

Prof Birhane said, "Studies and subsequent interventions would be most effectively focused on impacts on the marginalized populations who are disproportionately affected by both fabrications and negative stereotypes in model outputs. Additionally resources are needed for the education of the public, policymakers, and interdisciplinary scientists to give realistically informed views of how generative AI models work and to correct existing misinformation and hype surrounding these new technologies."

More information: Celeste Kidd et al, How generative AI models can distort human beliefs, *Science* (2023). DOI: 10.1126/science.adi0248. www.science.org/doi/10.1126/science.adi0248

Provided by Trinity College Dublin

Citation: Generative AI models are encoding biases and negative stereotypes in their users, say researchers (2023, June 22) retrieved 11 May 2024 from https://techxplore.com/news/2023-06-generative-ai-encoding-biases-negative.html