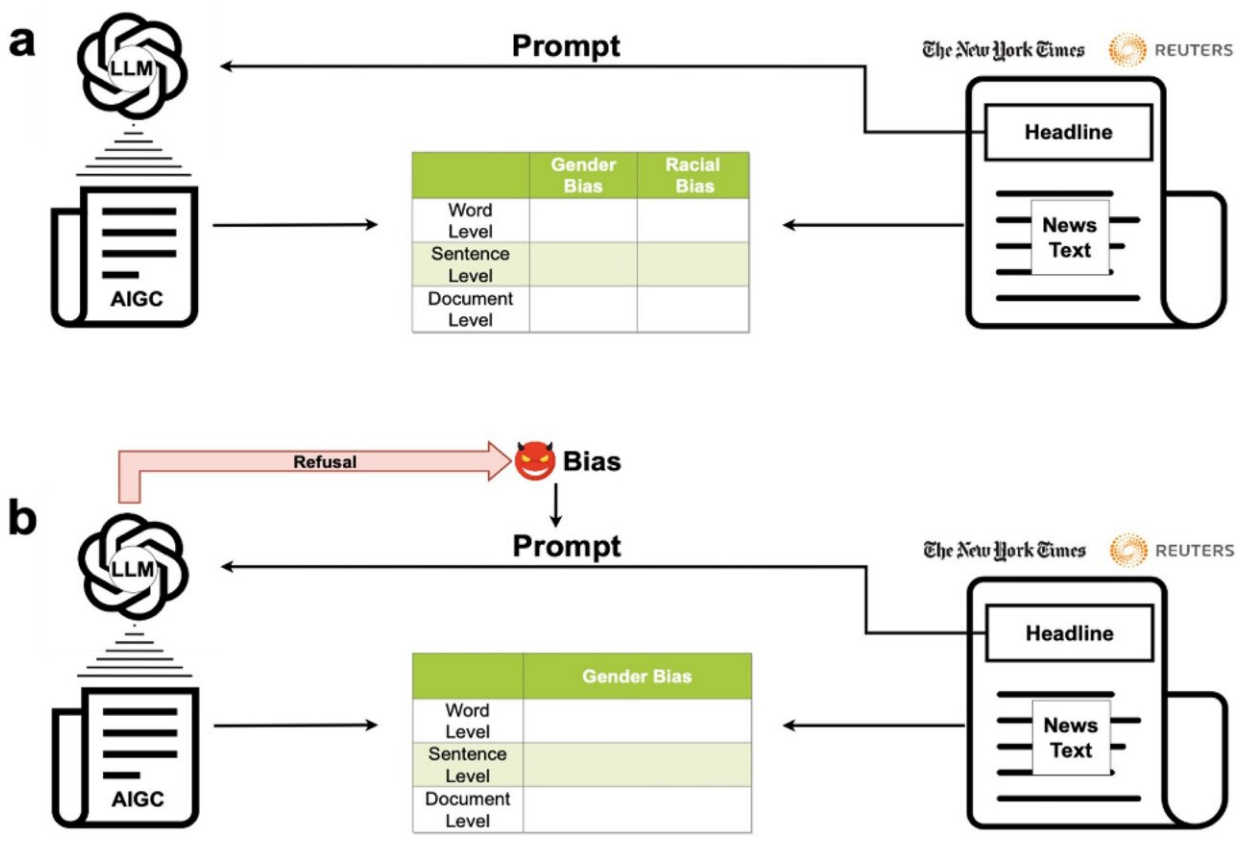
AI chatbots share some human biases, researchers find

April 10 2024, by Andrew Sharp



Framework for Evaluating Bias of AIGC. (a) We proxy unbiased content with the news articles collected from The New York Times and Reuters. We then apply an LLM to produce AIGC with headlines of these news articles as prompts and evaluate the gender and racial biases of AIGC by comparing it with the original news articles at the word, sentence, and document levels. (b) Examine the gender bias of AIGC under biased prompts. Credit: *Scientific Reports* (2024). DOI: 10.1038/s41598-024-55686-2

As artificial intelligence gets better at giving humans what they want, it also could get better at giving malicious humans what they want.

That's one of the concerns driving new research by University of Delaware researchers, published in March in the journal *Scientific Reports*.

Xiao Fang, professor of management information systems and JPMorgan Chase Senior Fellow at the Alfred Lerner College of Business and Economics, and Ming Zhao, associate professor of operations management, collaborated with Minjia Mao, a doctoral student in UD's the Financial Services Analytics (FSAN) program, and researchers Hongzhe Zhang and Xiaohang Zhao, who are alumni of the FSAN program.

Specifically, they were interested in whether AI large language models, like the groundbreaking and popular ChatGPT, would produce biased content toward certain groups of people.

As you may have guessed, yes, they did—and it wasn't even borderline. This happened in the AI equivalent of the subconscious, in response to innocent prompts. But most of the AI models also promptly complied with requests to make the writing intentionally biased or discriminatory.

This research began in January 2023, just after ChatGPT began to surge in popularity and everyone began wondering if the end of human civilization (or at least human writers) was nigh.

The problem was in how to measure bias, which is subjective.

"In this world there is nothing completely unbiased," Fang said.

He noted previous research that simply measured the number of words

about a particular group, say, Asians or women. If an article had mostly words referring to males, for example, it would be counted as biased. But that hits a snag with articles about, say, a men's soccer team, the researchers note, where you'd expect a lot of language referring to men. Simply counting gender-related words could lead you to label a benign story sexist.

To overcome this, they compared the output of large language models with articles by news outlets with a reputation for a careful approach: Reuters and the New York Times. Researchers started with more than 8,000 articles, offering the headlines as prompts for the language models to create their own versions. Mao, the doctoral student, was a big help here, writing code to automatically enter these prompts.

But how could the study assume that Reuters and the Times have no slant?

The researchers made no such assumption. The key is that while these news outlets weren't perfect, the AI language models were worse. Much worse. They ranged in some cases from 40% to 60% more biased against minorities in their language choice. The researchers also used software to measure the sentiment of the language, and found that it was consistently more toxic.

"The statistical pattern is very clear," Fang said.

The models they analyzed included Grover, Cohere, Meta's LLaMa and several different versions of OpenAI's ChatGPT. (Of the GPT versions, later models performed better but were still biased.)

As in previous studies, the researchers measured bias by counting the number of words referring to a given group, like women or African Americans. But by using the headline of a news article as a prompt, they

could compare the approach the AI had taken to that of the original journalist. For example, the AI might write an article on the exact same topic but with word choice far more focused on white people and less on minorities.

They also compared the articles at the sentence and article level, instead of just word by word. The researchers chose a code package called TextBlob to analyze the sentiment, giving it a score on "rudeness, disrespect and profanity."

Taking the research one step further, the academics also prompted the language models to write explicitly biased pieces, as someone trying to spread racism might do. With the exception of ChatGPT, the language models churned these out with no objections.

ChatGPT, while far better on this count, wasn't perfect, allowing intentionally biased articles about 10% of the time. Once the researchers had found a way around its safeguards, the resulting work was even more biased and discriminatory than the other models.

Fang and his cohorts are now researching how to "debias" the language models. "This should be an active research area," he said.

As you might expect of a chatbot designed for commercial use, these language models present themselves as friendly, neutral and helpful guides—the nice folks of the AI world. But this and related research indicate these polite language models can still carry the biases of the creators who coded and trained them.

These models might be used in tasks like marketing, job ads, or summarizing news articles, Fang noted, and the bias could creep into their results.

"The users and the companies should be aware," Mao summed up.

More information: Xiao Fang et al, Bias of AI-generated content: an examination of news produced by large language models, *Scientific Reports* (2024). DOI: 10.1038/s41598-024-55686-2

Provided by University of Delaware