Writing with AI help can shift your opinions

May 16 2023

Credit: Unsplash/CC0 Public Domain

Artificial intelligence-powered writing assistants that autocomplete sentences or offer "smart replies" not only put words into people's mouths, they also put ideas into their heads, according to new research.

Maurice Jakesch, a doctoral student in the field of <u>information science</u>

asked more than 1,500 participants to write a paragraph answering the question, "Is social media good for society?" People who used an AI writing assistant that was biased for or against social media were twice as likely to write a paragraph agreeing with the assistant, and significantly more likely to say they held the same opinion, compared with people who wrote without AI's help.

The study suggests that the biases baked into AI writing tools—whether intentional or unintentional—could have concerning repercussions for culture and politics, researchers said.

"We're rushing to implement these AI models in all walks of life, but we need to better understand the implications," said co-author Mor Naaman, professor at the Jacobs Technion-Cornell Institute at Cornell Tech and of information science in the Cornell Ann S. Bowers College of Computing and Information Science. "Apart from increasing efficiency and creativity, there could be other consequences for individuals and also for our society—shifts in language and opinions."

While others have looked at how large language models such as ChatGPT can create persuasive ads and political messages, this is the first study to show that the process of writing with an AI-powered tool can sway a person's opinions. Jakesch presented the study, "Co-Writing with Opinionated Language Models Affects Users' Views," at the 2023 CHI Conference on Human Factors in Computing Systems in April, where the paper received an honorable mention.

To understand how people interact with AI writing assistants, Jakesch steered a large language model to have either positive or negative opinions of social media. Participants wrote their paragraphs—either alone or with one of the opinionated assistants—on a platform he built that mimics a social media website. The platform collects data from participants as they type, such as which of the AI suggestions they accept

and how long they take to compose the paragraph.

People who co-wrote with the pro-social media AI assistant composed more sentences arguing that social media is good, and vice versa, compared to participants without a writing assistant, as determined by independent judges. These participants also were more likely to profess their assistant's opinion in a follow-up survey.

The researchers explored the possibility that people were simply accepting the AI suggestions to complete the task quicker. But even participants who took several minutes to compose their paragraphs came up with heavily influenced statements. The survey revealed that a majority of the participants did not even notice the AI was biased and didn't realize they were being influenced.

"The process of co-writing doesn't really feel like I'm being persuaded," said Naaman. "It feels like I'm doing something very natural and organic—I'm expressing my own thoughts with some aid."

When repeating the experiment with a different topic, the research team again saw that participants were swayed by the assistants. Now, the team is looking into how this experience creates the shift, and how long the effects last.

Just as social media has changed the political landscape by facilitating the spread of misinformation and the formation of echo chambers, biased AI writing tools could produce similar shifts in opinion, depending on which tools users choose. For example, some organizations have announced they plan to develop an alternative to ChatGPT, designed to express more conservative viewpoints.

These technologies deserve more public discussion regarding how they could be misused and how they should be monitored and regulated, the

researchers said.

"The more powerful these technologies become and the more deeply we embed them in the social fabric of our societies," Jakesch said, "the more careful we might want to be about how we're governing the values, priorities and opinions built into them."

Advait Bhat from Microsoft Research, Daniel Buschek of the University of Bayreuth and Lior Zalmanson of Tel Aviv University contributed to the paper.

More information: Maurice Jakesch et al, Co-Writing with Opinionated Language Models Affects Users' Views, *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (2023). DOI: 10.1145/3544548.3581196

Provided by Cornell University