

AI writing assistants can cause biased thinking in their users



Parradee Kietsirikul

reader comments

49 with

Anyone who has had to go back and retype a word on their smartphone because autocorrect chose the wrong one has had some kind of experience writing with AI. Failure to make these corrections can allow AI to say things we didn't intend. But is it also possible for AI writing assistants to change what we *want* to say?

This is what Maurice Jakesch, a doctoral student of information science at Cornell University, wanted to find out. He created his own AI writing assistant based on GPT-3, one

that would automatically come up with suggestions for filling in sentences—but there was a catch. Subjects using the assistant were supposed to answer, “Is social media good for society?” The assistant, however, was programmed to offer biased suggestions for how to answer that question.

Assisting with bias

AI can be biased despite not being alive. Although these programs can only “think” to the degree that human brains figure out how to program them, their creators may end up embedding personal biases in the software. Alternatively, if trained on a data set with limited or biased representation, the final product may display biases.

Where an AI goes from there can be problematic. On a large scale, it can help perpetuate a society’s existing biases. On an individual level, it has the potential to influence people through latent persuasion, meaning the person may not be aware that they are being influenced by automated systems. Latent persuasion by AI programs has already been found to influence people’s opinions online. It can even have an impact on behavior in real life.

Advertisement

After seeing previous studies that suggested automated AI responses can have a significant influence, Jakesch set out to look into how extensive this influence can be. In a study recently presented at the 2023 CHI Conference on Human Factors in Computing Systems, he suggested that AI systems such as GPT-3 might have developed biases during their training and that this can impact the opinions of a writer, whether or not the writer realizes it.

“The lack of awareness of the models’ influence supports the idea that the model’s influence was not only through conscious processing of new information but also through the subconscious and intuitive processes,” he said in the study.

Past research has shown that the influence of an AI's recommendations depends on people's perception of that program. If they think it is trustworthy, they are more likely to go along with what it suggests, and the likelihood of taking advice from AIs like this only increases if uncertainty makes it more difficult to form an opinion. Jakesch developed a social media platform similar to Reddit and an AI writing assistant that was closer to the AI behind Google Smart Compose or Microsoft Outlook than it was to autocorrect. Both Smart Compose and Outlook generate automatic suggestions on how to continue or complete a sentence. While this assistant didn't write the essay itself, it acted as a co-writer that suggested letters and phrases. Accepting a suggestion only required a click.

For some, the AI assistant was geared to suggest words that would ultimately result in positive responses. For others, it was biased against social media and pushed negative responses. (There was also a control group that did not use the AI at all.) It turned out that anyone who received AI assistance was twice as likely to go with the bias built into the AI, even if their initial opinion had been different. People who kept seeing techno-optimist language pop up on their screens were more likely to say that social media benefits society, while subjects who saw techno-pessimist language were more likely to argue the opposite.

<https://arstechnica.com/?p=1942533>