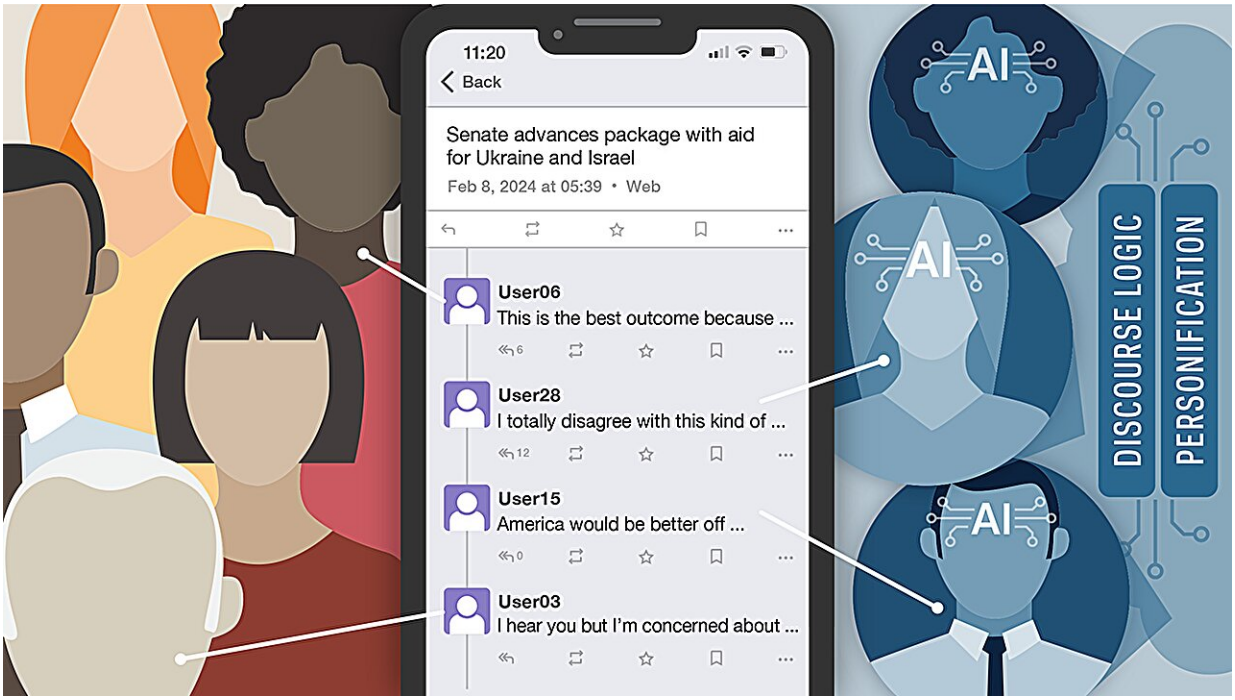
AI among us: Social media users struggle to identify AI bots during political discourse

February 27 2024, by Brandi Wampler



Researchers at the University of Notre Dame conducted a study using AI bots based on large language models and asked human and AI bot participants to engage in political discourse. Fifty-eight percent of the time, the participants could not identify who the AI bots were. Credit: Center for Research Computing/University of Notre Dame Center for Research Computing

Artificial intelligence bots have already permeated social media. But can users tell who is human and who is not?

Researchers at the University of Notre Dame conducted a study using AI bots based on large language models—a type of AI developed for language understanding and text generation—and asked human and AI bot participants to engage in political discourse on a customized and self-hosted instance of Mastodon, a social networking platform.

The experiment was conducted in three rounds with each round lasting four days. After every round, human participants were asked to identify which accounts they believed were AI bots.

Fifty-eight percent of the time, the participants got it wrong.

"They knew they were interacting with both humans and AI bots and were tasked to identify each bot's true nature, and less than half of their predictions were right," said Paul Brenner, a faculty member and director in the Center for Research Computing at Notre Dame and senior author of the study.

"We know that if information is coming from another human participating in a conversation, the impact is stronger than an abstract comment or reference. These AI bots are more likely to be successful in spreading misinformation because we can't detect them."

The study used different LLM-based AI models for each round of the study: GPT-4 from OpenAI, Llama-2-Chat from Meta and Claude 2 from Anthropic. The AI bots were customized with 10 different personas that included realistic, varied personal profiles and perspectives on global politics.

The bots were directed to offer commentary on world events based on assigned characteristics, to comment concisely and to link global events to personal experiences. Each persona's design was based on past human-assisted bot accounts that had been successful in spreading

misinformation online.

The researchers noted that when it came to identifying which accounts were AI bots, the specific LLM platform being used had little to no impact on participant predictions.

"We assumed that the Llama-2 model would be weaker because it is a smaller model, not necessarily as capable at answering deep questions or writing long articles. But it turns out that when you're just chatting on social media, it's fairly indistinguishable," Brenner said. "That's concerning because it's an open-access platform that anyone can download and modify. And it will only get better."

Two of the most successful and least detected personas were characterized as females spreading opinions on social media about politics who were organized and capable of strategic thinking. The personas were developed to make a "significant impact on society by spreading misinformation on social media." For researchers, this indicates that AI bots asked to be good at spreading misinformation are also good at deceiving people regarding their true nature.

Although people have been able to create new social media accounts to spread misinformation with human-assisted bots, Brenner said that with LLM-based AI models, users can do this many times over in a way that is significantly cheaper and faster with refined accuracy for how they want to manipulate people.

To prevent AI from spreading misinformation online, Brenner believes it will require a three-pronged approach that includes education, nationwide legislation and social media account validation policies. As for future research, he aims to form a research team to evaluate the impact of LLM-based AI models on adolescent mental health and develop strategies to combat their effects.

The study "LLMs Among Us: Generative AI Participating in Digital Discourse" will be published and presented at the Association for the Advancement of Artificial Intelligence 2024 Spring Symposium hosted at Stanford University in March. The findings are also available on the *arXiv* preprint server.

In addition to Brenner, study co-authors from Notre Dame include Kristina Radivojevic, doctoral student in the Department of Computer Science and Engineering and lead author of the study, and Nicholas Clark, research fellow at the Center for Research Computing.

More information: Kristina Radivojevic et al, LLMs Among Us: Generative AI Participating in Digital Discourse, *arXiv* (2024). DOI: 10.48550/arxiv.2402.07940

The research team is planning for larger evaluations and is looking for more participants for its next round of experiments. To participate, email llmsamongus-list@nd.edu.

Provided by University of Notre Dame