PHYS.ORG

Lawmakers struggle to differentiate AI and human emails

March 22 2023, by Kate Blackwood

Credit: Unsplash/CC0 Public Domain

Natural language models such as ChatGPT and GPT-4 open new opportunities for malicious actors to influence representative democracy, new Cornell research suggests.

A field experiment investigating how the natural language model GPT-3, the predecessor to the most recently released model, might be used to generate constituent email messages showed that legislators were only slightly less likely to respond to AI-generated messages (15.4%) than human-generated (17.3%).

The 2% difference, gleaned from more than 32,000 messages sent to about 7,000 state legislators in the U.S., was statistically significant but substantially small, the researchers said. The results highlight the potential threats this technology presents for democratic representation, but also suggest ways legislators might guard against AI-sourced astroturfing, the disingenuous practice of creating a sense of grassroot support, in this case by sending large volumes of content sympathetic to a particular issue.

The study, "The Potential Impact of Emerging Technologies on Democratic Representation: Evidence from a Field Experiment," co-authored by Sarah Kreps, the John L. Wetherill Professor in the Department of Government in the College of Arts and Sciences (A&S), director of the Cornell Jeb E. Brooks School Tech Policy Institute and adjunct professor of law, and Douglas Kriner, the Clinton Rossiter Professor in American Institutions in the Department of Government (A&S) and professor in the Brooks School, published March 20 in *New Media and Society*.

In recent years, new communication technologies have interfered with the democratic process multiple times, Kreps said. In the 2016 U.S. presidential election, Russian agents used micro-targeted social media advertisements to manipulate American voters and influence the outcome. And in 2017, the Federal Communications Commission's public comment lines were flooded with millions of messages generated by natural language models in response to a proposed rollback of regulations.

With these in mind, Kreps, who was an early academic collaborator of OpenAI, the organization that developed GPT-2, -3 and -4, and the more mainstream ChatGPT, wondered what malicious actors could do with more powerful language models now widely available.

"Could they generate misinformation or politically motivated, targeted content at scale?" she asked. "Could they effectively distort the democratic process? Or might they be able to generate large volumes of emails that seem like they're coming from constituents and thereby shift the legislative agenda toward the interests of a foreign government?"

In their experiment, conducted throughout 2020, Kreps and Kriner chose six current issues: reproductive rights, policing, tax levels, gun control, health policy and education policy. To create the human-generated messages, undergraduates associated with the student-run Cornell Political Union drafted emails to state legislators on each topic, advocating for the right-wing or left-wing position.

Then they produced machine generated constituency letters using GPT-3, training the system on 12 letters (a right and a left position for each of the six issues). They generated 100 different outputs for each of the ideologies and topics.

Many legislators and their staff did not dismiss the machine-generated content as inauthentic, the researchers said, as shown in the small difference in responses between AI and human content across the six issues.

Moreover, messages on gun control and health policy received virtually identical response rates, and on education policy, the response rate was higher for AI-generated messages, suggesting that "on these issues GPT-3 succeeded in producing content that was almost indistinguishable in the eyes of state legislative offices from human content," they wrote.

In feedback after the experiment, state legislators shared how they pick out fake emails, such as lack of geographical markers. Some said they represent districts so small they can spot fakes simply by looking at a name.

"It was heartening to hear that a lot of these legislators really understand their constituents and their voices, and that these AI-generated messages did not sound at all like something their constituents would write," Kreps said.

However, such local clues to authenticity would be more difficult for national-level senators and representatives to spot, the researchers said.

Technological tools employing the same type of neural networks can help differentiate real messages from fake, "but so can a discerning eye and digital literacy," Kreps said. "Legislators need to be trained to know what to look for."

As the capacity for electronic astroturfing increases, legislators may have to rely more heavily on other sources of information about constituency preferences, Kriner said, including district polling data and in-person events: "They travel around their constituencies holding town meetings and get a direct earful, at least from those most animated about an issue."

Both scholars are optimistic that democracy in America will survive this new threat.

"You could argue that the move from sitting down and writing a letter to writing an email was a lot bigger than between boilerplate online email templates and GPT-3," Kreps said. "We've adapted before, and democratic institutions will do it again."

More information: Sarah Kreps et al, The potential impact of emerging technologies on democratic representation: Evidence from a field experiment, *New Media & Society* (2023). DOI: 10.1177/14614448231160526

Provided by Cornell University

Citation: Lawmakers struggle to differentiate AI and human emails (2023, March 22) retrieved 28 April 2024 from https://phys.org/news/2023-03-lawmakers-struggle-differentiate-ai-human.html