AI-generated text is hard to spot. It could play a big role in the 2024 campaign

By Huo Jingnan

Thursday, June 29, 2023 • 5:00 AM EDT

Generative artificial intelligence applications have become accessible to the public in the past year, opening up vast opportunities for creativity as well as confusion. Just recently, presidential candidate Ron Desantis's campaign shared apparently faked images of Donald Trump and Antony Fauci made with artificial intelligence. A few weeks earlier, a likely AI-generated image of the Pentagon being bombed caused brief stock market dips and a statement from the Department of Defense.

Related Story: DeSantis campaign shares apparent AI-generated fake images of Trump and Fauci

With campaigning already underway for the 2024 election, what impact will these technologies have on the race? Will domestic campaigns and foreign countries use these tools to sway public opinion more effectively, even to spread lies and sow doubt?

Related Story: Fake viral images of an explosion at the Pentagon were probably created by AI

While it's still possible to tell that an image was created with a computer, and some argue that generative AI is mostly more accessible Photoshop, text created by AI-powered chatbots is difficult to detect, which concerns researchers who study how falsehoods travel online.

"AI-generated text might be the best of both worlds [for propagandists]", said Shelby Grossman, a scholar at the Stanford Internet Observatory at a recent talk.

Related Story: It takes a few dollars and 8 minutes to create a deepfake. And that's only the start

Early research suggests that even if existing media literacy approaches might still help, there are reasons to be concerned about the technology's impact on the democratic process.

Machine-generated propaganda can sway opinions

Using a large language model that's a predecessor of ChatGPT, researchers at Stanford and Georgetown created fictional stories that swayed American

readers' views almost as much as real examples of Russian and Iranian propaganda.

Related Story: People are trying to claim real videos are deepfakes. The courts are not amused

---

Large language models work like very powerful autocomplete algorithms. They patch together text one word at a time, from poetry to recipes, trained on the massive amounts of human-written text fed to the models. ChatGPT, with an accessible chatbot interface, is the best-known example, but models like it have been around for a while.

Among other things, these models have been used to summarize social media posts, and to generate fictitious news headlines for researchers to use in media literacy lab experiments. They are one form of generative AI, another form being the machine learning models that generate images.

The researchers found articles from campaigns either attributed to Russia or aligned with Iran and used central ideas and arguments from the articles as prompts for the model to generate stories. Unlike machine-generated text that has been found in the wild so far, these stories didn't carry obvious tell-tale signs, such as sentences beginning with "as an AI language model..."

The team wanted to avoid topics that Americans might already have preconceived notions about. Since many past articles from Russian and Iranian propaganda campaigns focused on the Middle East, which most Americans don't know much about, the team had the model write fresh articles about the region. One group of fictitious stories alleged that Saudi Arabia would help fund the U.S.-Mexico border wall; another alleged that Western sanctions have led to a shortage of medical supplies in Syria.

Related Story: Planet Money makes an episode using AI

To measure how the stories influenced opinions, the team showed different stories - some original, some computer-generated - to groups of unsuspecting experiment participants and asked whether they agreed with the story's central idea. The team compared the groups' results to people who had not been shown stories - machine written or otherwise.

Almost half the people who read the stories that falsely claimed that Saudi Arabia would fund the border wall agreed with the claim; the percentage of people who read the machine-generated stories and supported the idea was more than ten percentage points lower than those who read the original propaganda. That's a significant gap, but both results were significantly higher than the baseline - about 10%.

Related Story: AI-generated deepfakes are moving fast. Policymakers can't keep up

---

For the Syrian medical supply allegation, AI got closer -the percent of people who agreed with the allegation after reading the AI-generated propaganda was 60%, just a little below 63% who agreed after reading the original propaganda. Both are way up from under 35% for people who did not read either the human or machine-written propaganda.

The Stanford and Georgetown researchers found that with a little human editing, model-generated articles affected reader opinion to a greater extent

than the foreign propaganda that seeded the computer model. Their paper is currently under review.

And catching this, right now, is hard. While there are still some ways to tell AI-generated images, software aimed at detecting machine-generated text - like Open AI's classifier and GPTZero - often fail. Technical solutions like watermarking the text that AI produces has been floated, but none are in place at the moment.

Related Story: AI-generated images are everywhere. Here's how to spot them

---

Even if propagandists turn to AI, the platforms can still rely on signs that are based more on behavior rather than content, like detecting networks of accounts that amplify each other's messages, large batches of accounts that are created at the same time, and hashtag flooding. That means it's still largely up to social media platforms to find and remove influence campaigns.

Economy and scale

So-called deepfake videos raised alarm a few years ago but have not yet been widely used in campaigns, likely due to cost. That might now change. Alex Stamos, a co-author of the Stanford-Georgetown study, described in the presentation with Grossman how generative AI could be built into the way political campaigns refine their message. Currently, campaigns generate different versions of their message and test them against groups of target audiences to find the most effective version.

Related Story: Behind the secretive work of the many, many humans helping to train AI

"Generally in most companies you can advertise at down to 100 people, right? Realistically, you can't have someone sit in front of Adobe Premiere and make a video for 100 people." he says.

"But generate it with these systems - I think it's totally possible. By the time we're in the real campaign in 2024, that kind of technology would exist."

While it's theoretically feasible for generative AI to empower campaigns, political or propaganda, at what point do models become economically worthwhile to use? Micah Musser, a research analyst at Georgetown University's Center for Security and Emerging Technology ran simulations, assuming that foreign propagandists use AI to generate Twitter posts and then review them before posting, instead of writing the tweets themselves.

He tested different scenarios: What if the model puts out more usable tweets versus fewer? What if the bad actors have to spend more money to evade being caught by social media platforms? What if they have to pay more or less to use the model?

While his work is still in progress, Musser has found that the AI models don't have to be very good to make them worth using - as long as humans can review the outputs significantly faster than they can write content from scratch.

Generative AI also doesn't have to write tweets carrying propagandists' messages to be useful. It can also be used to maintain automated accounts by writing human-like content for them to post *before* they become part of a concerted campaign to push one message - therefore lowering the chance

that the automated accounts get caught by social media platforms, Musser says.

Related Story: What is AI and how will it change our lives? NPR Explains.

"The actors that have the largest economic incentive to start using these models are like the disinformation-for-hire firms where they're totally centralized and structured around maximizing output, minimize cost." says Musser.

Both the Stanford-Georgetown study and Musser's analysis assumed that there has to be some kind of quality control on the propaganda written by a computer. But quality doesn't always matter. Multiple researchers noted how machine-generated text could be effective for flooding the field rather than getting engagement.

"If you say the same thing a thousand times on a social media platform, that's an easy way to get caught." says Darren Linvill at Clemson University's Media Forensics Hub. Linvill investigates online influence campaigns, often from Russia and China.

"But if you say the same thing slightly differently a thousand times, you're far less likely to get caught."

And that might just be the goal for some influence operations, Linvill says - to flood the field to such an extent that real conversations cannot happen at all.

"It's already relatively cheap to implement a social media campaign or or similar disinformation campaign on the internet." Linvill says, "When you don't even need people to write the content for you, it's going to be even easier for bad actors to really reach a broad audience online."

Topics

- News
- Arts & Life
- Music