THE WALL STREET JOURNAL.

https://www.wsj.com/articles/readers-beware-ai-has-learned-to-create-fake-news-stories-11571018640

JOURNAL REPORTS: TECHNOLOGY

Readers Beware: AI Has Learned to Create Fake News Stories

Researchers warn about the risks of computer-generated articles—and release tools that ferret out fakes

By Asa Fitch

Updated Oct. 13, 2019 10:59 pm ET

Real-sounding but made-up news articles have become much easier to produce thanks to a handful of new tools powered by artificial intelligence—raising concerns about potential misuse of the technology.

What deepfakes did for video—producing clips of famous people appearing to say and do things they never said or did—these tools could do for news, tricking people into thinking the earth is flat, global warming is a hoax or a political candidate committed a crime when he or she didn't. While false articles are nothing new, these AI tools allow them to be generated in seconds by computer.

As far as experts know, the technology has been implemented only by researchers, and it hasn't been used maliciously. What's more, it has limitations that keep the stories from seeming *too* believable.

But many of the researchers who developed the technology, and people who have studied it, fear that as such tools get more advanced, they could spread misinformation or advance a political agenda. That's why some are sounding the alarm about the risks of computer-generated articles—and releasing tools that let people ferret out potentially fake stories.

"The danger is when there is already a lot of similar propaganda written by humans from which these neural language models can learn to generate similar articles," says Yejin Choi, an associate professor at the University of Washington, a researcher at the Allen Institute for Artificial Intelligence and part of a team that developed a fake-news tool. "The quality of such neural fake news can look quite convincing to humans."

Stop the presses

The first entry in a powerful new generation of synthetic-text tools was unveiled in February, when OpenAI, a San Francisco-based research body backed by prominent tech names like LinkedIn co-founder Reid Hoffman, launched the GPT-2. The software produces genuine-sounding news articles—as well as other types of passages, from fiction to conversations—by drawing on its analysis of 40 gigabytes of text across eight million webpages. Researchers developed the OpenAI software because they knew powerful speech-generation would eventually appear in the wild and wanted to handle its release responsibly.

The GPT-2 system worked so well that in an August survey of 500 people, a majority found its synthetic articles credible. In one group of participants, 72% found a GPT-2 article credible, compared with 83% who found a genuine article credible.

"Large-scale synthesized disinformation is not only possible but is cheap and credible," says Sarah Kreps, a professor at Cornell University who co-wrote the research. Its spread across the internet, she says, could open the way for malicious influence campaigns. Even if people *don't* believe the fake articles are accurate, she says, the knowledge that such stories are out there could have a damaging effect, eroding people's trust in the media and government.

JOURNAL REPORT

- Read more at WSJ.com/AIreport

MORE IN ARTIFICIAL INTELLIGENCE

- What Will an AI World Be Like?
- Relationship Help From an Algorithm
- Debate: Should You Be Able to Sell Your Personal Data?

Given the potential risks associated with giving the world full access to the GPT-2, OpenAI decided not to release it immediately, instead putting out a more limited version for researchers to study and potentially develop tools that could detect artificially generated texts in the wild.

In the months that followed, other researchers replicated OpenAI's work. In June, Dr. Choi and her colleagues at the University of Washington and the Allen Institute for Artificial Intelligence posted a tool on the institute's website called Grover, positioning it as a piece of software that could both generate convincing false news stories and use the same technology to detect others' artificial news by ferreting out telltale textual patterns.

Then, in August, Israel's AI21 Labs put a language-generation tool called HAIM on its website. It asserted on its site that risks of releasing text-generation tools into the wild were overblown, and that there were beneficial uses of such automatically generated texts, including simplifying and speeding the writing process.

The human touch

Yoav Shoham, co-founder of AI21, said in an interview that the effectiveness of these text-generation tools as propaganda machines was limited because they can't incorporate political context well enough to score points with target audiences. Even if an AI can produce a real-looking article, Mr. Shoham said, a machine can't grasp, say, the dynamics of a feud between two politicians and craft a false story that discredits one of them in a nuanced way.

"They have the appearance of making sense, but they don't," Mr. Shoham said.

Plus, very often articles go off on strange tangents for reasons the researchers don't completely understand—the systems are often black boxes, generating text based on their own analyses of existing documents.

Ultimately, Dr. Choi says, producing effective propaganda requires machines to have a broader understanding of how the world works and a fine-tuned sense of how to target such material, something only a human overseeing the process could bring to the table.

"Fine-grained control of the content is not within the currently available technology," she says.

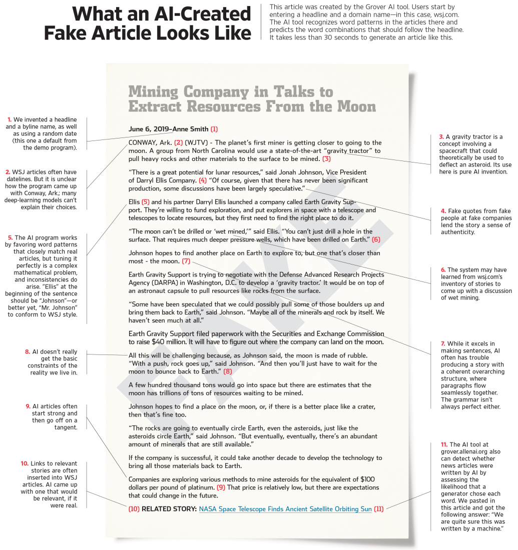
While so far it doesn't appear that any of the technology has been used as propaganda, the threat is real enough that the U.S. Defense Department's Defense Advanced Research Projects Agency, or Darpa, in late August unveiled a program called Semantic Forensics. The project aims to defend against a wide range of automated disinformation attacks, including text-based ones.

Private groups are also developing systems to detect fake stories. Along with the freely available online tool Grover, researchers at the Massachusetts Institute of Technology and Harvard introduced a text inspector (http://gltr.io/dist/index.html) in March. The software uses similar techniques as Grover, predicting whether a passage is AI-made by taking a chunk of text and analyzing how likely a language-generation model would be to pick the word that actually appears next.

But if language-generation models change how they select words and phrases in the future, detection won't necessarily improve at the same rate, says Jack Clark, OpenAI's policy director. Ever more complex language-generation systems are proliferating rapidly, driven by researchers and developers who are training new models on larger pools of data. OpenAI already has a model trained on more than 1.5 billion parameters that it hasn't yet released to the public.

“Increasingly large language models could feasibly either naturally develop or be trained to better approximate human patterns of writing as they get bigger,” Mr. Clark says.

Mr. Fitch is a Wall Street Journal reporter in San Francisco. He can be reached at asa.fitch@wsj.com.



What an AI-Created Fake Article Looks Like

This article was created by the Grover AI tool. Users start by entering a headline and a domain name—in this case, wsj.com. The AI tool recognizes word patterns in the articles there and predicts the word combinations that should follow the headline. It takes less than 30 seconds to generate an article like this.

Mining Company in Talks to Extract Resources From the Moon

June 6, 2019–Anne Smith (1)

CONWAY, Ark. (2) (WJTV) - The planet’s first miner is getting closer to going to the moon. A group from North Carolina would use a state-of-the-art “gravity tractor” to pull heavy rocks and other materials to the surface to be mined. (3)

“There is a great potential for lunar resources,” said Jonah Johnson, Vice President of Darryl Ellis Company. (4) “Of course, given that there has never been significant production, some discussions have been largely speculative.”

Ellis (5) and his partner Darryl Ellis launched a company called Earth Gravity Support. They’re willing to fund exploration, and put explorers in space with a telescope and telescopes to locate resources, but they first need to find the right place to do it.

“The moon can’t be drilled or ‘wet mined,’” said Ellis. “You can’t just drill a hole in the surface. That requires much deeper pressure wells, which have been drilled on Earth.” (6)

Johnson hopes to find another place on Earth to explore to, but one that’s closer than most - the moon. (7)

Earth Gravity Support is trying to negotiate with the Defense Advanced Research Projects Agency (DARPA) in Washington, D.C. to develop a ‘gravity tractor.’ It would be on top of an astronaut capsule to pull resources like rocks from the surface.

“Some have been speculated that we could possibly pull some of those boulders up and bring them back to Earth,” said Johnson. “Maybe all of the minerals and rock by itself. We haven’t seen much at all.”

Earth Gravity Support filed paperwork with the Securities and Exchange Commission to raise $40 million. It will have to figure out where the company can land on the moon.

All this will be challenging because, as Johnson said, the moon is made of rubble. “With a push, rock goes up,” said Johnson. “And then you’ll just have to wait for the moon to bounce back to Earth.” (8)

A few hundred thousand tons would go into space but there are estimates that the moon has trillions of tons of resources waiting to be mined.

Johnson hopes to find a place on the moon, or, if there is a better place like a crater, then that’s fine too.

“The rocks are going to eventually circle Earth, even the asteroids, just like the asteroids circle Earth,” said Johnson. “But eventually, eventually, there’s an abundant amount of minerals that are still available.”

If the company is successful, it could take another decade to develop the technology to bring all those materials back to Earth.

Companies are exploring various methods to mine asteroids for the equivalent of $100 dollars per pound of platinum. (9) That price is relatively low, but there are expectations that could change in the future.

(10) **RELATED STORY:** NASA Space Telescope Finds Ancient Satellite Orbiting Sun (11)

1. We invented a headline and a byline name, as well as using a random date (this one a default from the demo program).

2. WSJ articles often have datelines. But it is unclear how the program came up with Conway, Ark.; many deep-learning models can’t explain their choices.

3. A gravity tractor is a concept involving a spacecraft that could theoretically be used to deflect an asteroid. Its use here is pure AI invention.

4. Fake quotes from fake people at fake companies lend the story a sense of authenticity.

5. The AI program works by favoring word patterns that closely match real articles, but tuning it perfectly is a complex mathematical problem, and inconsistencies do arise. ”Ellis” at the beginning of the sentence should be “Johnson”—or better yet, “Mr. Johnson” to conform to WSJ style.

6. The system may have learned from wsj.com’s inventory of stories to come up with a discussion of wet mining.

7. While it excels in making sentences, AI often has trouble producing a story with a coherent overarching structure, where paragraphs flow seamlessly together. The grammar isn’t always perfect either.

8. AI doesn’t really get the basic constraints of the reality we live in.

9. AI articles often start strong and then go off on a tangent.

10. Links to relevant stories are often inserted into WSJ articles. AI came up with one that would be relevant, if it were real.

11. The AI tool at grover.allenai.org also can detect whether news articles were written by AI by assessing the likelihood that a generator chose each word. We pasted in this article and got the following answer: “We are quite sure this was written by a machine.”

SHARE YOUR THOUGHTS

What consequences do you foresee with this technology? Join the conversation below.