The Guardian

New AI fake text generator may be too dangerous to release, say creators

The Elon Musk-backed nonprofit company OpenAI declines to release research publicly for fear of misuse

The AI wrote a new passage of fiction set in China after being fed the opening line of Nineteen Eighty-Four by George Orwell (pictured). Photograph: Mondadori/Getty Images

Alex Hern

Thu 14 Feb 2019 17.00 GMT

The creators of a revolutionary AI system that can write news stories and works of fiction – dubbed "deepfakes for text" – have taken the unusual step of not releasing their research publicly, for fear of potential misuse.

OpenAI, an nonprofit research company backed by Elon Musk, Reid Hoffman, Sam Altman, and others, says its new AI model, called GPT2 is so good and the risk of malicious use so high that it is breaking from its normal practice of releasing the full research to the public in order to allow more time to discuss the ramifications of the technological breakthrough.

At its core, GPT2 is a text generator. The AI system is fed text, anything from a few words to a whole page, and asked to write the next few sentences based on its

predictions of what should come next. The system is pushing the boundaries of what was thought possible, both in terms of the quality of the output, and the wide variety of potential uses.

When used to simply generate new text, GPT2 is capable of writing plausible passages that match what it is given in both style and subject. It rarely shows any of the quirks that mark out previous AI systems, such as forgetting what it is writing about midway through a paragraph, or mangling the syntax of long sentences.

Feed it the opening line of George Orwell's Nineteen Eighty-Four - "It was a bright cold day in April, and the clocks were striking thirteen" - and the system recognises the vaguely futuristic tone and the novelistic style, and continues with:

"I was in my car on my way to a new job in Seattle. I put the gas in, put the key in, and then I let it run. I just imagined what the day would be like. A hundred years from now. In 2045, I was a teacher in some school in a poor part of rural China. I started with Chinese history and history of science."

Feed it the first few paragraphs of a Guardian story about Brexit, and its output is plausible newspaper prose, replete with "quotes" from Jeremy Corbyn, mentions of the Irish border, and answers from the prime minister's spokesman.

One such, completely artificial, paragraph reads: "Asked to clarify the reports, a spokesman for May said: 'The PM has made it absolutely clear her intention is to leave the EU as quickly as is possible and that will be under her negotiating mandate as confirmed in the Queen's speech last week.'"

From a research standpoint, GPT2 is groundbreaking in two ways. One is its size, says Dario Amodei, OpenAI's research director. The models "were 12 times bigger, and the dataset was 15 times bigger and much broader" than the previous state-of-the-art AI model. It was trained on a dataset containing about 10m articles, selected by trawling the social news site Reddit for links with more than three votes. The vast collection of text weighed in at 40 GB, enough to store about 35,000 copies of Moby Dick.

The amount of data GPT2 was trained on directly affected its quality, giving it more knowledge of how to understand written text. It also led to the second breakthrough. GPT2 is far more general purpose than previous text models. By structuring the text that is input, it can perform tasks including translation and summarisation, and pass simple reading comprehension tests, often performing as well or better than other AIs that have been built specifically for those tasks.

That quality, however, has also led OpenAI to go against its remit of pushing AI forward and keep GPT2 behind closed doors for the immediate future while it assesses what malicious users might be able to do with it. "We need to perform experimentation to find out what they can and can't do," said Jack Clark, the charity's head of policy. "If you can't anticipate all the abilities of a model, you have to prod it to see what it can do. There are many more people than us who are better at thinking what it can do maliciously."

To show what that means, OpenAI made one version of GPT2 with a few modest tweaks that can be used to generate infinite positive - or negative - reviews of products. Spam

and fake news are two other obvious potential downsides, as is the AI's unfiltered nature . As it is trained on the internet, it is not hard to encourage it to generate bigoted text, conspiracy theories and so on.

Instead, the goal is to show what is possible to prepare the world for what will be mainstream in a year or two's time. "I have a term for this. The escalator from hell," Clark said. "It's always bringing the technology down in cost and down in price. The rules by which you can control technology have fundamentally changed.

"We're not saying we know the right thing to do here, we're not laying down the line and saying 'this is the way' ... We are trying to develop more rigorous thinking here. We're trying to build the road as we travel across it."

Topics

- Artificial intelligence (AI)
- Elon Musk
- Computing
- Journalism books
- George Orwell