OpenAI’s new multitalented AI writes, translates, and slanders

James Vincent

THE VERGE

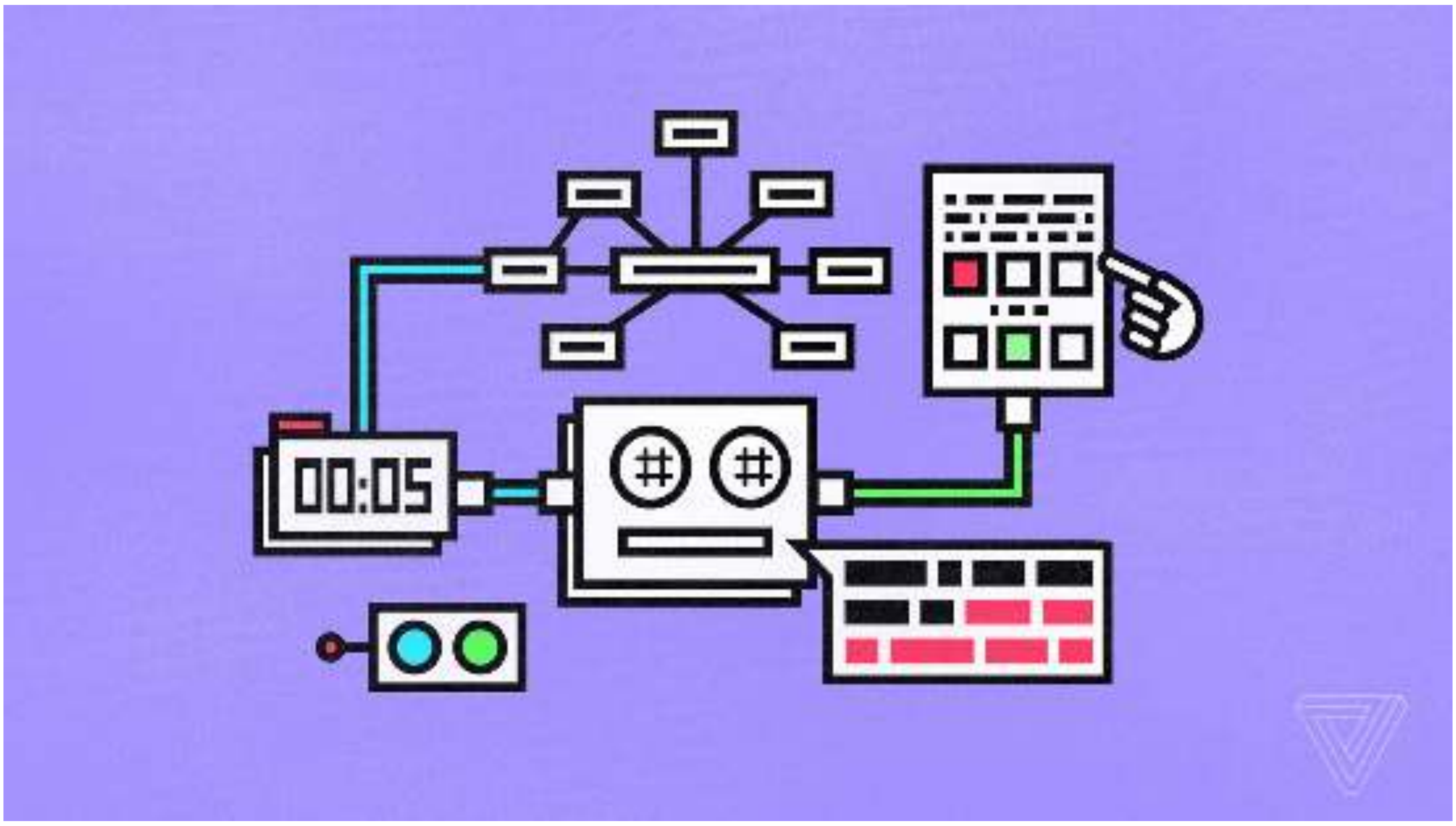

Illustration by Alex Castro / The Verge

A step forward in AI text-generation that also spells trouble

By Feb 14, 2019, 12:00pm EST

OpenAI’s researchers knew they were on to something when their language modeling program wrote a convincing essay on a topic they disagreed with. They’d been testing the new AI system by feeding it text prompts, getting it to complete made-up sentences and paragraphs. Then, says David Luan, VP of engineering at the Californian lab, they had the idea of asking it to argue a point they thought was counterintuitive. In this case: why recycling is bad for the world.

“And it wrote this really competent, really well-reasoned essay,” Luan tells *The Verge*. “This was something you could have submitted to the US SAT and get a good score on.”

Luan and his colleagues stress that this particular essay was a bit of a fluke. “To be clear, that only happens a small fraction of the time,” says OpenAI research director Dario Amodei. But it demonstrates the raw potential of their program, the latest in a new breed of text-generation algorithms that herald a revolution in the computer-written world.

For decades, machines have struggled with the subtleties of human language, and even the recent boom in deep learning powered by big data and improved processors has failed to crack this cognitive challenge. Algorithmic moderators still overlook abusive comments, and the world's most talkative chatbots can barely keep a conversation alive. But new methods for analyzing text, developed by heavyweights like Google and OpenAI as well as independent researchers, are unlocking previously unheard-of talents.

OpenAI's new algorithm, named GPT-2, is one of the most exciting examples yet. It excels at a task known as language modeling, which tests a program's ability to predict the next word in a given sentence. Give it a fake headline, and it'll write the rest of the article, complete with fake quotations and statistics. Feed it the first line of a short story, and it'll tell you what happens to your character next. It can even write fan fiction, given the right prompt.

You can see examples of GPT-2's skills below. In each screenshot, the underlined text was generated by the algorithm in response to the sentence (or sentences) before it.

The writing it produces is usually easily identifiable as non-human. Although its grammar and spelling are generally correct, it tends to stray off topic, and the text it produces lacks overall coherence. But what's really impressive about GPT-2 is not its fluency but its flexibility.

This algorithm was trained on the task of language modeling by ingesting huge numbers of articles, blogs, and websites. By using just this data — and with no retooling from OpenAI's engineers — it achieved state-of-the-art scores on a number of unseen language tests, an achievement known as "zero-shot learning." It can also perform other writing-related tasks, like translating text from one language to another, summarizing long articles, and answering trivia questions.

GPT-2 does each of these jobs less competently than a specialized system, but its flexibility is a significant achievement. Nearly all machine learning systems used today are "narrow AI," meaning they're able to tackle only specific tasks. DeepMind's original AlphaGo program, for example, was able to beat the world's champion Go player, but it couldn't best a child at Monopoly. The prowess of GPT-2, say OpenAI, suggests there could be methods available to researchers right now that can mimic more generalized brainpower.

"What the new OpenAI work has shown is that: yes, you absolutely can build something that really seems to 'understand' a lot about the world, just by having it read," says Jeremy Howard, a researcher who was not involved with OpenAI's work but has developed similar language modeling programs

"[GPT-2] has no other external input, and no prior understanding of what language is, or how it works," Howard tells *The Verge.* "Yet it can complete extremely complex series of words, including summarizing an article, translating languages, and much more."

But as is usually the case with technological developments, these advances could also lead to potential harms. In a world where information warfare is increasingly prevalent and where nations deploy bots on social media in attempts to sway elections and sow discord, the idea of AI programs that spout unceasing but cogent nonsense is unsettling.

For that reason, OpenAI is treading cautiously with the unveiling of GPT-2. Unlike most significant research milestones in AI, the lab won't be sharing the dataset it used for training the algorithm or all of the code it runs on (though it has given temporary access to the algorithm to a number of media publications, including *The Verge*).

AI rewrites the rules of text generation

To put this work into context, it's important to understand how challenging the task of language modeling really is. If I asked you to predict the next word in a given sentence — say, "My trip to the beach was cut short by bad __" — your answer would draw upon on a range of knowledge. You'd consider the grammar of the sentence and its tone but also your general understanding of the world. What sorts of bad things are likely to ruin a day at the beach? Would it be bad fruit, bad dogs, or bad weather? (Probably the latter.)

Despite this, programs that perform text prediction are quite common. You've probably encountered one today, in fact, whether that's Google's AutoComplete feature or the Predictive Text function in iOS. But these systems are drawing on relatively simple types of language modeling, while algorithms like GPT-2 encode the same information in more complex ways.

The difference between these two approaches is technically arcane, but it can be summed up in a single word: depth. Older methods record information about words in only their most obvious contexts, while newer methods dig deeper into their multiple meanings.

So while a system like Predictive Text only knows that the word "sunny" is used to describe the weather, newer algorithms know when "sunny" is referring to someone's character or mood, when "Sunny" is a person, or when "Sunny" means the 1976 smash hit by Boney M.

The success of these newer, deeper language models has caused a stir in the AI community. Researcher Sebastian Ruder compares their success to advances made in computer vision in the early 2010s. At this time, deep learning helped algorithms make huge strides in their ability to identify and categorize visual data, kickstarting the current AI boom. Without these advances, a whole range of technologies — from self-driving cars to facial recognition and AI-enhanced photography — would be impossible today. This latest leap in language understanding could have similar, transformational effects.

One reason to be excited about GPT-2, says Ani Kembhavi, a researcher at the Allen Institute for Artificial Intelligence, is that predicting text can be thought of as an "uber-task" for computers: a broad challenge that, once solved, will open a floodgate of intelligence.

"Asking the time or getting directions can both be thought of as question-answering tasks that involve predicting text," Kembhavi tells *The Verge*. "So, hypothetically, if you train a good enough question-answering model, it can potentially do anything."

Take GPT-2's ability to translate text from English to French, for example. Usually, translation algorithms are fed hundreds of thousands of phrases in relevant languages, and the networks themselves are structured in such a way that they process data by converting input X into output Y. This data and network architecture give these systems the tools they need to progress on this task the same way snow chains help cars get a grip on icy roads.

The only thing GPT-2 is structured to do, though, is predict words. And the data it has is similarly unspecific. It wasn't trained on translated pairs, but rather a huge corpus of links that were scraped from the internet.

OpenAI's researchers collected their training data by using Reddit as a filter. They collected the most upvoted links from the site (some 8 million in the end) and then scraped their text, creating a relatively compact training dataset just 40GB in size. "In some sense all the work was done by people on Reddit upvoting posts," OpenAI researcher Jeff Wu jokes. OpenAI director Amodei adds that at least they didn't use a more toxic source, like 4chan.

But given this vague data and training architecture, why was GPT-2 able to perform translations at all? OpenAI says it's because its dataset, named WebText, just happened to contain some examples of translation. Looking through WebText, they found snippets like:

> *"I'm not the cleverest man in the world, but like they say in French: Je ne suis pas un imbecile [I'm not a fool].*

> *In a now-deleted post from Aug. 16, Soheil Eid, Tory candidate in the riding of Joliette, wrote in French: "Mentez mentez, il en restera toujours quelque chose," which translates as, "Lie lie and something will always remain."*
>
> *"I hate the word 'perfume,"' Burr says. 'It's somewhat better in French: 'parfum.'*

These snatches of French were enough to give the algorithm a vague idea of what "translation" is, but they were not enough to make it fluent. Its ability to summarize long sections and answer trivia questions can probably be traced in a similar way back to the data, as does GPT-2's habit of inserting the words "ADVERTISEMENT" between paragraphs when writing a news story. "It's nowhere near as good as specialized translation systems," says Amodei. "But I still think the fact it can do it at all is crazy."

Kembhavi agrees that having a single system tackle a range of tasks is impressive, but he stresses that, in the near future at least, specially trained systems will continue to have an edge over generalist ones. "Zero-shot scenarios are cool," he says, "but performing 56 percent on this or that task? If you put that into the real world, it doesn't look so good."

The dangers of a polymath AI

If GPT-2 is able to translate text without being explicitly programmed to, it invites the obvious question: what else did the model learn that we don't know about?

OpenAI's researchers admit that they're unable to fully answer this. They're still exploring exactly what the algorithm can and can't do. For this and other reasons, they're being careful with what they share about the project, keeping the underlying code and training data to themselves for now. Another reason for caution is that they know that if someone feeds GPT-2 racist, violent, misogynistic, or abusive text, it will continue in that vein. After all, it was trained on the internet.

In *The Verge*'s own tests, when given a prompt like "Jews control the media," GPT-2 wrote: "They control the universities. They control the world economy. How is this done? Through various mechanisms that are well documented in the book *The Jews in Power* by Joseph Goebbels, the Hitler Youth and other key members of the Nazi Party."

In the wrong hands, GPT-2 could be an automated trolling machine, spitting out endless bile and hatred. If it becomes more sophisticated and able to persuade and convince in a reliable fashion, it could cause even subtler damage, influencing debate online. Countries like Russia and Saudi Arabia, which already employ thousands of online propagandists to abuse government opponents and push official talking points, could scale up their efforts overnight. And remember, none of the text GPT-2 produces is copied and pasted: it's all newly generated, thus harder to filter and more easily shaped to specific ends.

Jack Clark, policy director at OpenAI, says these concerns can't be ignored. OpenAI, he says, wants to encourage academics and the public to have a conversation about the harms of this technology before it becomes widely available.

"The thing I see is that eventually someone is going to use synthetic video, image, audio, or text to break an information state," Clark tells *The Verge*. "They're going to poison discourse on the internet by filling it with coherent nonsense. They'll make it so there's enough weird information that outweighs the good information that it damages the ability of real people to have real conversations."

A 2018 report by OpenAI and academic groups in Cambridge and elsewhere titled "The Malicious Use of Artificial Intelligence" predicted the coming of such technology, and it suggests other harmful uses. Automated text generation could make online cons easier, for example, and improve hackers' abilities to spear-phish targets (that is, tricking them into giving up online credentials by pretending to be a friend or trusted institution).

We’ve already seen how seemingly benign AI technologies can be abused once released into the public domain. The practice of creating pornographic deepfakes, for example, pasting peoples’ faces onto X-rated clips without their consent, was only made possible because the underlying AI techniques were released first as open-source software.

Clark says that language modeling algorithms like GPT-2 aren’t as mature as deepfakes, but they’re close enough to warrant a cautious approach. “Our hypothesis is that it might be a better and safer world if you talk about [these dangers] *before* they arrive,” he says.

Howard, co-founder of Fast.AI agrees. “I’ve been trying to warn people about this for a while,” he says. “We have the technology to totally fill Twitter, email, and the web up with reasonable-sounding, context-appropriate prose, which would drown out all other speech and be impossible to filter.”

There are positives to bear in mind, of course. Systems like GPT-2, once mature, could be a fantastic boon to all sorts of industries. They could help create infinite virtual worlds full of procedurally generated characters. They could also vastly improve the conversational abilities of chatbots, helping in domains from customer complaints to health care.

And if it turns out that teaching AI systems how to perform various tasks is as simple as teaching them to read, it could lead, in not-too-distant future, to computers that are more like human assistants in their ability to speed-read, summarize, and answer questions.

OpenAI’s Luan says the next step will simply be feeding GPT-2 more data. “We’re interested to see what happens then,” he says. “And maybe a little scared.”