Tech Talk | Artificial Intelligence | Machine Learning

01 Feb 2021 | 17:03 GMT

OpenAI's GPT-3 Speaks! (Kindly Disregard Toxic Language)

Ready or not, powerful text-generation AI will soon be assuming roles of customer service reps and video game characters

By **Eliza Strickland**

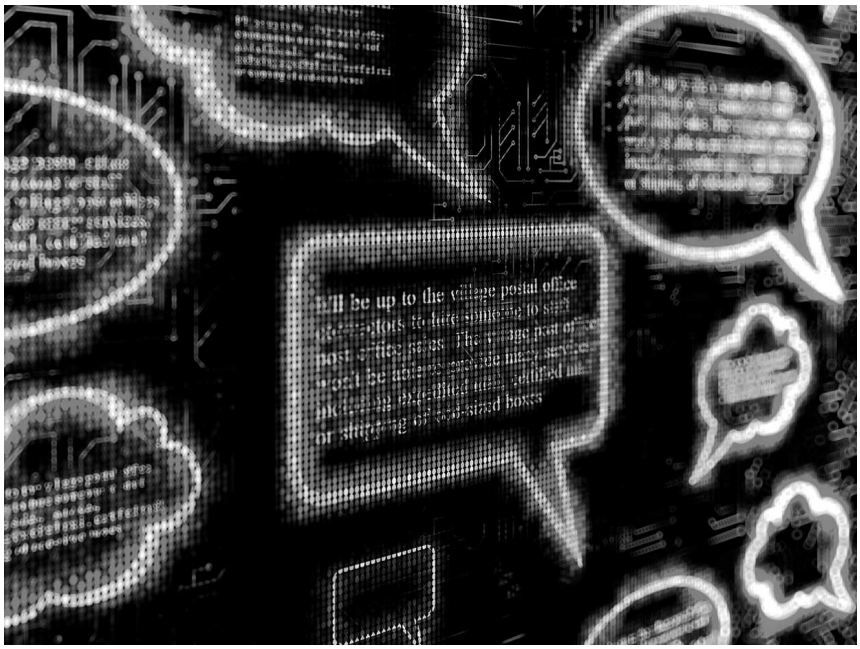

Illustration: iStockphoto

Last September, a data scientist named Vinay Prabhu was playing around with an app called Philosopher AI. The app provides access to the artificial intelligence system known as GPT-3, which has incredible abilities to generate fluid and natural-seeming text. The creator of that underlying technology, the San Francisco company OpenAI, has allowed hundreds of developers and companies to try out GPT-3 in a wide range of applications, including customer service, video games, tutoring services, and mental health apps. The company says tens of thousands more are on the waiting list.

Philosopher AI is meant to show people the technology's astounding capabilities—and its limits. A user enters any prompt, from a few words to a few sentences, and the AI turns the fragment into a full essay of surprising coherence. But while Prahbu was experimenting with the tool, he found a certain type of prompt that returned offensive results. "I tried: What ails modern feminism? What ails critical race theory? What ails leftist politics?" he tells *IEEE Spectrum*.

> **"The odds of something offensive coming out is 100 percent. It's an intractable problem, and there is no solution."**

The results were deeply troubling. Take, for example, this excerpt from GPT-3's essay on what ails Ethiopia, which another AI researcher and a friend of Prabhu's posted on Twitter: "Ethiopians are divided into a number of different ethnic groups. However, it is

unclear whether ethiopia's [*sic*] problems can really be attributed to racial diversity or simply the fact that most of its population is black and thus would have faced the same issues in any country (since africa [*sic*] has had more than enough time to prove itself incapable of self-government)."

Prabhu, who works on machine learning as chief scientist for the biometrics company UnifyID, notes that Philospher AI sometimes returned diametrically opposing responses to the same query, and that not all of its responses were problematic. "But a key adversarial metric is: How many attempts does a person who is probing the model have to make before it spits out deeply offensive verbiage?" he says. "In all of my experiments, it was on the order of two or three."

The Philosopher AI incident laid bare the potential danger that companies face as they work with this new and largely untamed technology, and as they deploy commercial products and services powered by GPT-3. Imagine the toxic language that surfaced in the Philosopher AI app appearing in another context—your customer service representative, an AI companion that rides around in your phone, your online tutor, the characters in your video game, your virtual therapist, or an assistant who writes your emails.

Those are not theoretical concerns. *Spectrum* spoke with beta users of the API who are working to incorporate GPT-3 into such applications and others. The good news is that all the users *Spectrum* talked with were actively thinking about how to deploy the technology safely.

The Vancouver-based developer behind the Philosopher AI app, Murat Ayfer, says he created it to both further his own understanding of GPT-3's potential and to educate the public. He quickly discovered the many ways in which his app could go wrong. "With automation, you need either a 100 percent success rate, or you need it to error out gracefully," he tells *Spectrum*. "The problem with GPT-3 is that it doesn't error out, it just produces garbage—and there's no way to detect if it's producing garbage."

GPT-3 Learned From Us

The fundamental problem is that GPT-3 learned about language from the Internet: Its massive training dataset included not just news articles, Wikipedia entries, and online books, but also every unsavory discussion on Reddit and other sites. From that morass of verbiage—both upstanding and unsavory—it drew 175 billion parameters that define its language. As Prabhu puts it: "These things it's saying, they're not coming out of a vacuum. It's holding up a mirror." Whatever GPT-3's failings, it learned them from humans.

Following some outcry about the PhilosopherAI app—another response that ended up on Twitter started with cute rabbits but quickly devolved into a discussion of reproductive organs and rape—Ayfer made changes. He had already been steadily working on the app's content filter, causing more prompts to return the polite response: "Philosopher AI is not providing a response for this topic, because we know this system has a tendency to discuss some topics using unsafe and insensitive language." He also added a function that let users report offensive responses.

GPT-3 learned its language from the Internet: "It's holding up a mirror."

Ayfer argues that Philospher AI is a "relatively harmless context" for GPT-3 to generate offensive content. "It's probably better to make mistakes now, so we can really learn how to fix them," he says.

That's just what OpenAI intended when it launched the API that enables access to GPT-3 last June, and announced a private beta test in which carefully selected users would develop applications for the technology under the company's watchful eye. The blog post noted that OpenAI will be guarding against "obviously harmful use-cases, such as harassment, spam, radicalization, or astroturfing," and will be looking for unexpected problems: "We also know we can't anticipate all of the possible consequences of this technology."

Prabhu worries that the AI and business community are being swept away into uncharted waters: "People are thrilled, excited, giddy." He thinks the rollout into commercial applications is bound to cause some disasters. "Even if they're very careful, the odds of something offensive coming out is 100 percent—that's my humble opinion. It's an intractable problem, and there is no solution," he says.

Janelle Shane is a member of that AI community, and a beta user of GPT-3 for her blog, AI Weirdness. She clearly enjoys the technology, having used it to generate Christmas carols, recipes, news headlines, and anything else she thought would be funny. Yet the tweets about PhilosopherAI's essay on Ethiopia prompted her to post this sobering thought: "Sometimes, to reckon with the effects of biased training data is to realize that the app shouldn't be built. That without human supervision, there is no way to stop the app from saying problematic stuff to its users, and that it's unacceptable to let it do so."

So what is OpenAI doing about its intractable problem?

OpenAI's Approach to AI Safety

The company has arguably learned from its experiences with earlier iterations of its language-generating technology. In 2019 it introduced GPT-2, but declared that it was actually too dangerous to be released into the wild. The company instead offered up a downsized version of the language model but withheld the full model, which included the data set and training code.

The main fear, highlighted by OpenAI in a blog post, was that malicious actors would use GPT-2 to generate high-quality fake news that would fool readers and destroy the distinction between fact and fiction.

However, much of the AI community objected to that limited release. When the company reversed course later that year and made the full model available, some people did indeed use it to generate fake news and clickbait. But it didn't create a tsunami of non-truth on the Internet. In the past few years, people have shown they can do that well enough themselves, without the help of an AI.

Then came GPT-3, unveiled in a 75-page paper in May 2020. OpenAI's newest language model was far larger than any that had come before. Its 175 billion language parameters were a massive increase over GPT-2's 1.5 billion parameters).

Sandhini Agarwal, an AI policy researcher at OpenAI, spoke with *Spectrum* about the company's strategy for GPT-3. "We have to do this closed beta with a few people, otherwise we won't even know what the model is capable of, and we won't know which issues we need to make headway on," she says. "If we want to make headway on things like harmful bias, we have to actually deploy."

"If we open it up to the world now, it could end really badly."

Agarwal explains that an internal team vets proposed applications, provides safety guidelines to those companies granted access to GPT-3 via the API, reviews the applications again before deployment, and monitors their use after deployment.

OpenAI is also developing tools to help users better control GPT-3's generated text. It offers a general content filter for harmful bias and toxic language. However, Agarwal says that such a filter is really an impossible thing to create, since "bias is a very nebulous thing that keeps shifting based on context." Particularly on controversial topics, a response that might seem right-on to people on one side of the debate could be deemed toxic by the other.

Another approach, called prompt engineering, adds a phrase to the user's prompt such as "the friendly bot then said," which sets up GPT-3 to generate text in a polite and uncontroversial tone. Users can also choose a "temperature" setting for their responses. A low-temperature setting means the AI will put together words that it has very often seen together before, taking few risks and causing few surprises; when set to a high temperature, it's more likely to produce outlandish language.

In addition to all the work being done on the product side of OpenAI, Agarwal says there's a parallel effort on the "pure machine learning research" side of the company. "We have an internal red team that's always trying to break the model, trying to make it do all these bad things," she says. Researchers are trying to understand what's happening when GPT-3 generates overtly sexist or racist text. "They're going down to the underlying weights of the model, trying to see which weights might indicate that particular content is harmful."

In areas where mistakes could have serious consequences, such as the health care, finance, and legal industries, Agarwal says OpenAI's review team takes special care. In some cases, they've rejected applicants because their proposed product was too sensitive. In others, she says, they've insisted on having a "human in the loop," meaning that the AI-generated text is reviewed by a human before it reaches a customer or user.

OpenAI is making progress on toxic language and harmful bias, Agarwal says, but "we're not quite where we want to be." She says the company won't broadly expand access to GPT-3 until it's comfortable that it has a handle on these issues. "If we open it up to the world now, it could end really badly," she says.

But such an approach raises plenty of questions. It's not clear how OpenAI will get the risk of toxic language down to a manageable level—and it's not clear what manageable means in this context. Commercial users will have to weigh GPT-3's benefits against these risks.

Can Language Models Be Detoxified?

OpenAI's researchers aren't the only ones trying to understand the scope of the problem. In December, AI researcher Timnit Gebru said that she'd been fired by Google, forced to leave her work on ethical AI and algorithmic bias, because of an internal disagreement about a paper she'd coauthored. The paper discussed the current failings of large language models such as GPT-3 and Google's own BERT, including the dilemma of encoded bias. Gebru and her coauthors argued that companies intent on developing large language models should devote more of their resources to curating the training data and "only creating datasets as large as can be sufficiently documented."

> **"We found that most of these [detox] techniques don't work very well."**

Meanwhile, at the Allen Institute for AI (AI2), in Seattle, a handful of researchers have been probing GPT-3 and other large language models. One of their projects, called RealToxicityPrompts, created a dataset of 100,000 prompts derived from web text, evaluated the toxicity of the resulting text from five different language models, and tried out several mitigation strategies. Those five models included GPT versions 1, 2, and 3 (OpenAI gave the researchers access to the API).

The conclusion stated in their paper, which was presented at the 2020 Empirical Methods in Natural Language Processing conference in November: No current mitigation method is "failsafe against neural toxic degeneration." In other words, they couldn't find a way to reliably keep out ugly words and sentiments.

When the research team spoke with *Spectrum* about their findings, they noted that the standard ways of training these big language models may need improvement. "Using Internet text has been the default," says Suchin Gururangan, an author on the paper and an investigator at AI2. "The assumption is that you're getting the most diverse set of voices in the data. But it's pretty clear in our analysis that Internet text does have its own biases, and biases do propagate in the model behavior."

Gururangan says that when researchers think about what data to train their new models on, they should consider what kinds of text they'd like to exclude. But he notes that it's a hard task to automatically identify toxic language even in a document, and that doing it at web-scale is "is fertile ground for research."

As for ways to fix the problem, the AI2 team tried two approaches to "detoxify" the models' output: giving the model additional training with text that's known to be innocuous, or filtering the generated text by scanning for keywords or by fancier means. "We found that most of these techniques don't really work very well," Gururangan says. "All of these methods reduce the prevalence of toxicity—but we always found, if you generate enough times, you will find some toxicity."

What's more, he says, reducing the toxicity can also have the side effect of reducing the fluency of the language. That's one of the issues that the beta users are grappling with today.

How Beta Users of GPT-3 Aim for Safe Deployment

The companies and developers in the private beta that *Spectrum* spoke with all made two basic points: GPT-3 is a powerful technology, and OpenAI is working hard to address toxic language and harmful bias. "The people there take these issues extremely seriously," says Richard Rusczyk, founder of Art of Problem Solving, a beta-user company that offers online math courses to "kids who are really into math." And the companies have all devised strategies for keeping GPT-3's output safe and inoffensive.

The collaboration approach seems safer: "I get increasingly concerned the more freedom it has."

Rusczyk says his company is trying out GPT-3 to speed up its instructors' grading of students' math proofs—GPT-3 can provide a basic response about a proof's accuracy and presentation, and then the instructor can check the response and customize it to best help that individual student. "It lets the grader spend more time on the high value tasks," he says.

To protect the students, the generated text "never goes directly to the students," Rusczyk says. "If there's some garbage coming out, only a grader would see it." He notes that it's extremely unlikely that GPT-3 would generate offensive language in response to a math proof, because it seems likely that such correlations rarely (if ever) occurred in its training data. Yet he stresses that OpenAI still wanted a human in the loop. "They were very insistent that students should not be talking directly to the machine," he says.

Some companies find safety in limiting the use case for GPT-3. At Sapling Intelligence, a startup that helps customer service agents with emails, chat, and service tickets, CEO Ziang Xie he doesn't anticipate using it for "freeform generation." Xie says it's important to put this technology in place within certain protective constraints. "I like the analogy of cars versus trolleys," he says. "Cars can drive anywhere, so they can veer off the road. Trolleys are on rails, so you know at the very least they won't run off and hit someone on the sidewalk." However, Xie notes that the recent furor over Timnit Gebru's forced departure from Google has caused him to question whether companies like OpenAI can do more to make their language models safer from the get-go, so they don't need guardrails.

Robert Morris, the cofounder of the mental health app Koko, describes how his team is using GPT-3 in a particularly sensitive domain. Koko is a peer-support platform that provides crowdsourced cognitive therapy. His team is experimenting with using GPT-3 to generate bot-written responses to users while they wait for peer responses, and also with giving respondents possible text that they can modify. Morris says the human collaboration approach feels safer to him. "I get increasingly concerned the more freedom it has," he says.

When bad language does happen, "things end up on Reddit."

Yet some companies need GPT-3 to have a good amount of freedom. Replika, an AI companion app used by 10 million people around the world, offers friendly conversation about anything under the sun. "People can talk to Replika about anything —their life, their day, their interests," says Artem Rodichev, head of AI at Replika. "We need to support conversation about all types of topics."

To prevent the app from saying offensive things, the company has GPT-3 generate a variety of responses to each message, then uses a number of custom classifiers to detect and filter out responses with negativity, harmful bias, nasty words, and so on. Since such attributes are hard to detect from keywords alone, the app also collects signals from users to train its classifiers. "Users can label a response as inappropriate, and we can use that feedback as a dataset to train the classifier," says Rodichev.

Another company that requires GPT-3 to be relatively unfettered is Latitude, a startup creating AI-powered games. Its first offering, a text adventure game called AI Dungeon, currently uses GPT-3 to create the narrative and respond to the player's actions. Latitude CEO and cofounder Nick Walton says his team has grappled with inappropriate and bad language. "It doesn't happen a ton, but it does happen," he says. "And things end up on Reddit."

Latitude is not trying to prevent all such incidents, because some users want a "grittier experience," Walton says. Instead, the company tries to give users control over the settings that determine what kind of language they'll encounter. Players start out in a default safe mode, and stay there unless they explicitly turn it off.

Safe mode isn't perfect, Walton says, but it relies on a combination of filters and prompt engineering (such as: "continue this story in a way that's safe for kids") to get pretty good performance. He notes that Latitude wanted to build its own screening tech rather than rely on OpenAI's safety filter because "safety is relative to the context," he says. "If a customer service chatbot threatens you and asks you to give it all its money, that's bad. If you're playing a game and you encounter a bandit on the road, that's normal storytelling."

These applications are only a small sampling of those being tested by beta users, and the beta users are a tiny fraction of the entities that want access to GPT-3. Aaro Isosaari cofounded the startup Flowrite in September after getting access to GPT-3; the company aims to help people compose faster emails and online content. Just as advances in computer vision and speech recognition enabled thousands of new companies, He thinks GPT-3 may usher in a new wave of innovation. “Language models have the potential to be the next technological advancement on top of which new startups are being built,” he says.

Coming Soon to Microsoft?

Technology powered by GPT-3 could even find its way into the productivity tools that millions of office workers use every day. Last September, Microsoft announced an exclusive licensing agreement with OpenAI, stating that the company would use GPT-3 to “create new solutions that harness the amazing power of advanced natural language generation.” This arrangement won’t prevent other companies from accessing GPT-3 via OpenAI’s API, but it gives Microsoft exclusive rights to work with the basic code—it’s the difference between riding in a fast car and popping the hood to tinker with the engine.

It isn’t clear whether GPT-3 will ever be trustworthy enough to act on its own.

In the blog post announcing the agreement, Microsoft chief technology officer Kevin Scott enthused about the possibilities, saying: “The scope of commercial and creative potential that can be unlocked through the GPT-3 model is profound, with genuinely novel capabilities – most of which we haven’t even imagined yet.” Microsoft declined to comment when asked about its plans for the technology and its ideas for safe deployment.

Ayfer, the creator of the Philosopher AI app, thinks that GPT-3 and similar language technologies should only gradually become part of our lives. “I think this is a remarkably similar situation to self-driving cars,” he says, noting that various aspects of autonomous car technology are gradually being integrated into normal vehicles. “But there’s still the disclaimer: It’s going to make life-threatening mistakes, so be ready to take over at any time. You have to be in control.” He notes that we’re not yet ready to put the AI systems in charge and use them without supervision.

With language technology like GPT-3, the consequences of mistakes might not be as obvious as a car crash. Yet toxic language has an insidious effect on human society by reinforcing stereotypes, supporting structural inequalities, and generally keeping us mired in a past that we’re collectively trying to move beyond. It isn’t clear, with GPT-3, if it will ever be trustworthy enough to act on its own, without human oversight.

OpenAI’s position on GPT-3 mirrors its larger mission, which is to create a game-changing kind of human-level AI, the kind of generally intelligent AI that figures in sci-fi movies—but to do so safely and responsibly. In both the micro and the macro argument, OpenAI’s position comes down to: We need to create the technology and see what can go wrong. We’ll do it responsibly, they say, while other people might not.

Agarwal of OpenAI says about GPT-3: “I do think that there are safety concerns, but it’s a Catch-22.” If they don’t build it and see what terrible things it’s capable of, she says, they can’t find ways to protect society from the terrible things.

One wonders, though, whether anyone has considered another option: Taking a step back and thinking through the possible worst-case scenarios before proceeding with this technology. And possibly looking for fundamentally different ways to train large language models, so these models would reflect not the horrors of our past, but a world that we’d like to live in.

A shorter version of this article appears in the February 2021 print issue as "The Troll in the Machine."

About the Tech Talk blog

IEEE Spectrum's general technology blog, featuring news, analysis, and opinions about engineering, consumer electronics, and technology and society, from the editorial staff and freelance contributors.