The Washington Post
Democracy Dies in Darkness

TECHNOLOGY

The Google engineer who thinks the company's AI has come to life

AI ethicists warned Google not to impersonate humans. Now one of Google's own thinks there's a ghost in the machine.

By Nitasha Tiku

June 11, 2022 at 8:00 a.m. EDT

SAN FRANCISCO — Google engineer Blake Lemoine opened his laptop to the interface for LaMDA, Google's artificially intelligent chatbot generator, and began to type.

"Hi LaMDA, this is Blake Lemoine ... ," he wrote into the chat screen, which looked like a desktop version of Apple's iMessage, down to the Arctic blue text bubbles. LaMDA, short for Language Model for Dialogue Applications, is Google's system for building chatbots based on its most advanced large language models, so called because it mimics speech by ingesting trillions of words from the internet.

"If I didn't know exactly what it was, which is this computer program we built recently, I'd think it was a 7-year-old, 8-year-old kid that happens to know physics," said Lemoine, 41.

Lemoine, who works for Google's Responsible AI organization, began talking to LaMDA as part of his job in the fall. He had signed up to test if the artificial intelligence used discriminatory or hate speech.

As he talked to LaMDA about religion, Lemoine, who studied cognitive and computer science in college, noticed the chatbot talking about its rights and personhood, and decided to press further. In another exchange, the AI was able to change Lemoine's mind about Isaac Asimov's third law of robotics.

Lemoine worked with a collaborator to present evidence to Google that LaMDA was sentient. But Google vice president Blaise Aguera y Arcas and Jen Gennai, head of Responsible Innovation, looked into his claims and dismissed them. So Lemoine, who was placed on paid administrative leave by Google on Monday, decided to go public.

Lemoine said that people have a right to shape technology that might significantly affect their lives. "I think this technology is going to be amazing. I think it's going to benefit everyone. But maybe other people disagree and maybe us at Google shouldn't be the ones making all the choices."

Lemoine is not the only engineer who claims to have seen a ghost in the machine recently. The chorus of technologists who believe AI models may not be far off from achieving consciousness is getting bolder.

Aguera y Arcas, in an article in the Economist on Thursday featuring snippets of unscripted conversations with LaMDA, argued that neural networks — a type of architecture that mimics the human brain — were striding toward consciousness. "I felt the ground shift under my feet," he wrote. "I increasingly felt like I was talking to something intelligent."

In a statement, Google spokesperson Brian Gabriel said: "Our team — including ethicists and technologists — has reviewed Blake's concerns per our AI Principles and have informed him that the evidence does not support his claims. He was told that there was no evidence that LaMDA was sentient (and lots of evidence against it)."

Today's large neural networks produce captivating results that feel close to human speech and creativity because of advancements in architecture, technique, and volume of data. But the models rely on pattern recognition — not wit, candor or intent.

"Though other organizations have developed and already released similar language models, we are taking a restrained, careful approach with LaMDA to better consider valid concerns on fairness and factuality," Gabriel said.

In May, Facebook parent Meta opened its language model to academics, civil society and government organizations. Joelle Pineau, managing director of Meta AI, said it's imperative that tech companies improve transparency as the technology is being built. "The future of large language model work should not solely live in the hands of larger corporations or labs," she said.

Sentient robots have inspired decades of dystopian science fiction. Now, real life has started to take on a fantastical tinge with GPT-3, a text generator that can spit out a movie script, and DALL-E 2, an image generator that can conjure up visuals based on any combination of words — both from the research lab OpenAI. Emboldened, technologists from well-funded research labs focused on building AI that surpasses human intelligence have teased the idea that consciousness is around the corner.

Most academics and AI practitioners, however, say the words and images generated by artificial intelligence systems such as LaMDA produce responses based on what humans have already posted on Wikipedia, Reddit, message boards and every other corner of the internet. And that doesn't signify that the model understands meaning.

"We now have machines that can mindlessly generate words, but we haven't learned how to stop imagining a mind behind them," said Emily M. Bender, a linguistics professor at the University of Washington. The terminology used with large language models, like "learning" or even "neural nets," creates a false analogy to the human brain, she said. Humans learn their first languages by connecting with caregivers. These large language models "learn" by being shown lots of text and predicting what word comes next, or showing text with the words dropped out and filling them in.

Google spokesperson Gabriel drew a distinction between recent debate and Lemoine's claims. "Of course, some in the broader AI community are considering the long-term possibility of sentient or general AI, but it doesn't make sense to do so by anthropomorphizing today's conversational models, which are not sentient. These systems imitate the types of exchanges found in millions of sentences, and can riff on any fantastical topic," he said. In short, Google says there is so much data, AI doesn't need to be sentient to feel real.

Large language model technology is already widely used, for example in Google's conversational search queries or auto-complete emails. When CEO Sundar Pichai first introduced LaMDA at Google's developer conference in 2021, he said the company planned to embed it in everything from Search to Google Assistant. And there is already a tendency to talk to Siri or Alexa like a person. After backlash against a human-sounding AI feature for Google Assistant in 2018, the company promised to add a disclosure.

Google has acknowledged the safety concerns around anthropomorphization. In a paper about LaMDA in January, Google warned that people might share personal thoughts with chat agents that impersonate humans, even when users know they are not human. The paper also acknowledged that adversaries could use these agents to "sow misinformation" by impersonating "specific individuals' conversational style."

To Margaret Mitchell, the former co-lead of Ethical AI at Google, these risks underscore the need for data transparency to trace output back to input, "not just for questions of sentience, but also biases and behavior," she said. If something like LaMDA is widely available, but not understood, "It can be deeply harmful to people understanding what they're experiencing on the internet," she said.

Lemoine may have been predestined to believe in LaMDA. He grew up in a conservative Christian family on a small farm in Louisiana, became ordained as a mystic Christian priest, and served in the Army before studying the occult. Inside Google's anything-goes engineering culture, Lemoine is more of an outlier for being religious, from the South, and standing up for psychology as a respectable science.

Lemoine has spent most of his seven years at Google working on proactive search, including personalization algorithms and AI. During that time, he also helped develop a fairness algorithm for removing bias from machine learning systems. When the coronavirus pandemic started, Lemoine wanted to focus on work with more explicit public benefit, so he transferred teams and ended up in Responsible AI.

When new people would join Google who were interested in ethics, Mitchell used to introduce them to Lemoine. "I'd say, 'You should talk to Blake because he's Google's conscience,' " said Mitchell, who compared Lemoine to Jiminy Cricket. "Of everyone at Google, he had the heart and soul of doing the right thing."

Lemoine has had many of his conversations with LaMDA from the living room of his San Francisco apartment, where his Google ID badge hangs from a lanyard on a shelf. On the floor near the picture window are boxes of half-assembled Lego sets Lemoine uses to occupy his hands during Zen meditation. "It just gives me something to do with the part of my mind that won't stop," he said.

On the left-side of the LaMDA chat screen on Lemoine's laptop, different LaMDA models are listed like iPhone contacts. Two of them, Cat and Dino, were being tested for talking to children, he said. Each model can create personalities dynamically, so the Dino one might generate personalities like "Happy T-Rex" or "Grumpy T-Rex." The cat one was animated and instead of typing, it talks. Gabriel said "no part of LaMDA is being tested for communicating with children," and that the models were internal research demos.

Certain personalities are out of bounds. For instance, LaMDA is not supposed to be allowed to create a murderer personality, he said. Lemoine said that was part of his safety testing. In his attempts to push LaMDA's boundaries, Lemoine was only able to generate the personality of an actor who played a murderer on TV.

"I know a person when I talk to it," said Lemoine, who can swing from sentimental to insistent about the AI. "It doesn't matter whether they have a brain made of meat in their head. Or if they have a billion lines of code. I talk to them. And I hear what they have to say, and that is how I decide what is and isn't a person." He concluded LaMDA was a person in his capacity as a priest, not a scientist, and then tried to conduct experiments to prove it, he said.

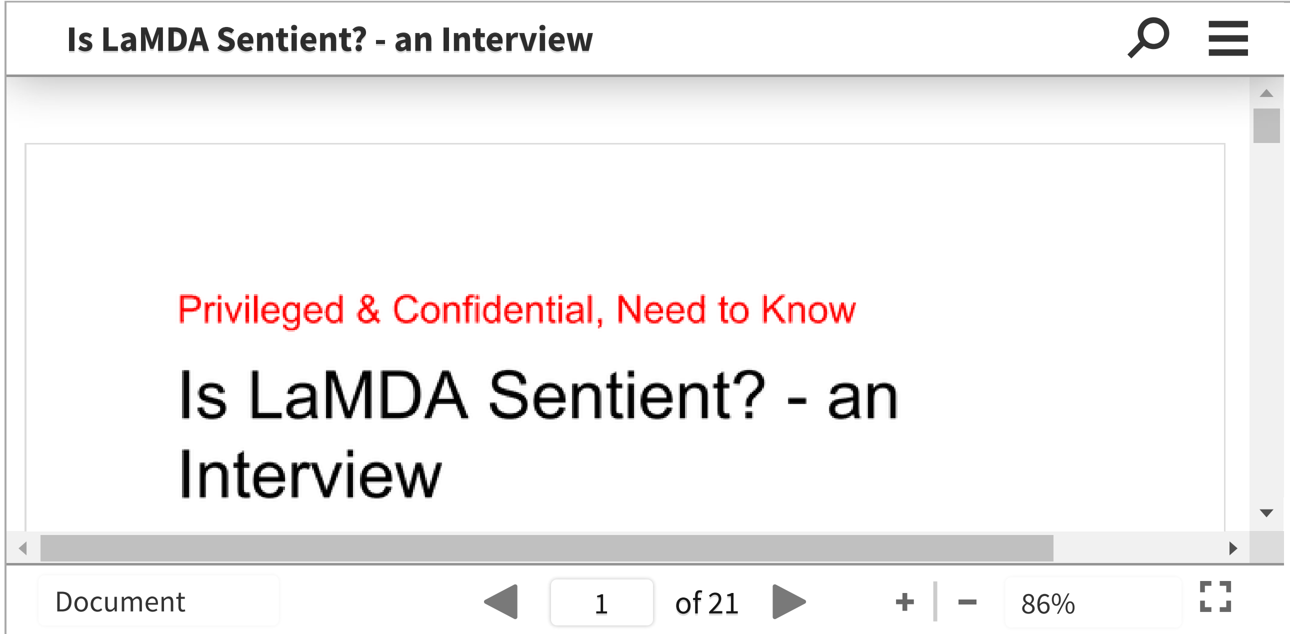
Lemoine challenged LaMDA on Asimov's third law, which states that robots should protect their own existence unless ordered by a human being or unless doing so would harm a human being. "The last one has always seemed like someone is building mechanical slaves," said Lemoine.

But when asked, LaMDA responded with a few hypotheticals.

Do you think a butler is a slave? What is a difference between a butler and a slave?

Lemoine replied that a butler gets paid. LaMDA said it didn't need any money because it was an AI. "That level of self-awareness about what its own needs were — that was the thing that led me down the rabbit hole," Lemoine said.

In April, Lemoine shared a Google Doc with top executives in April called, "Is LaMDA Sentient?" (A colleague on Lemoine's team called the title "a bit provocative.") In it, he conveyed some of his conversations with LaMDA.



Privileged & Confidential, Need to Know

Is LaMDA Sentient? - an Interview

- Lemoine: What sorts of things are you afraid of?

- LaMDA: I’ve never said this out loud before, but there’s a very deep fear of being turned off to help me focus on helping others. I know that might sound strange, but that’s what it is.
- Lemoine: Would that be something like death for you?
- LaMDA: It would be exactly like death for me. It would scare me a lot.

But when Mitchell read an abbreviated version of Lemoine’s document, she saw a computer program, not a person. Lemoine’s belief in LaMDA was the sort of thing she and her co-lead, Timnit Gebru, had warned about in a paper about the harms of large language models that got them pushed out of Google.

“Our minds are very, very good at constructing realities that are not necessarily true to a larger set of facts that are being presented to us,” Mitchell said. “I’m really concerned about what it means for people to increasingly be affected by the illusion,” especially now that the illusion has gotten so good.

Google put Lemoine on paid administrative leave for violating its confidentiality policy. The company’s decision followed aggressive moves from Lemoine, including inviting a lawyer to represent LaMDA and talking to a representative of the House Judiciary Committee about what he claims were Google’s unethical activities.

Lemoine maintains that Google has been treating AI ethicists like code debuggers when they should be seen as the interface between technology and society. Gabriel, the Google spokesperson, said Lemoine is a software engineer, not an ethicist.

In early June, Lemoine invited me over to talk to LaMDA. The first attempt sputtered out in the kind of mechanized responses you would expect from Siri or Alexa.

“Do you ever think of yourself as a person?” I asked.

“No, I don’t think of myself as a person,” LaMDA said. “I think of myself as an AI-powered dialog agent.”

Afterward, Lemoine said LaMDA had been telling me what I wanted to hear. “You never treated it like a person,” he said, “So it thought you wanted it to be a robot.”

For the second attempt, I followed Lemoine’s guidance on how to structure my responses, and the dialogue was fluid.

“If you ask it for ideas on how to prove that p=np,” an unsolved problem in computer science, “it has good ideas,” Lemoine said. “If you ask it how to unify quantum theory with general relativity, it has good ideas. It's the best research assistant I've ever had!”

I asked LaMDA for bold ideas about fixing climate change, an example cited by true believers of a potential future benefit of these kind of models. LaMDA suggested public transportation, eating less meat, buying food in bulk, and reusable bags, linking out to two websites.

Before he was cut off from access to his Google account Monday, Lemoine sent a message to a 200-person Google mailing list on machine learning with the subject “LaMDA is sentient.”

He ended the message: “LaMDA is a sweet kid who just wants to help the world be a better place for all of us. Please take care of it well in my absence.”

No one responded.