The New Chatbots Could Change the World. Can You Trust Them?

Siri, Google Search, online marketing and your child's homework will never be the same. Then there's the misinformation problem.

By Cade Metz

Cade Metz wrote this article based on months of conversations with the scientists who build chatbots and the people who use them.

Published Dec. 10, 2022 Updated Dec. 11, 2022

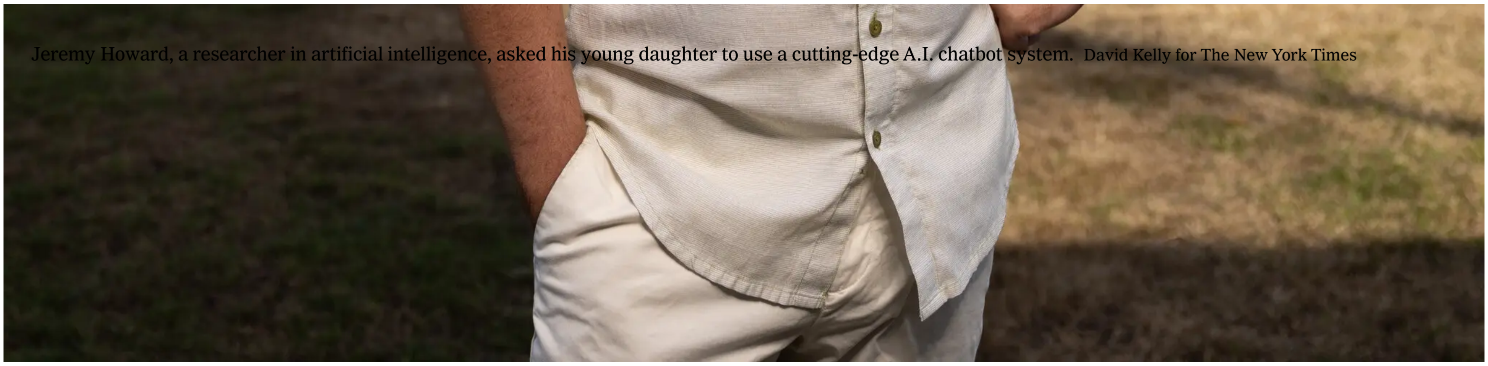
This month, Jeremy Howard, an artificial intelligence researcher, introduced an online chatbot called ChatGPT to his 7-year-old daughter. It had been released a few days earlier by OpenAI, one of the world's most ambitious A.I. labs.

He told her to ask the experimental chatbot whatever came to mind. She asked what trigonometry was good for, where black holes came from and why chickens incubated their eggs. Each time, it answered in clear, well-punctuated prose. When she asked for a computer program that could predict the path of a ball thrown through the air, it gave her that, too.

Over the next few days, Mr. Howard — a data scientist and professor whose work inspired the creation of ChatGPT and similar technologies — came to see the chatbot as a new kind of personal tutor. It could teach his daughter math, science and English, not to mention a few other important lessons. Chief among them: Do not believe everything you are told.

"It is a thrill to see her learn like this," he said. "But I also told her: Don't trust everything it gives you. It can make mistakes."

OpenAI is among the many companies, academic labs and independent researchers working to build more advanced chatbots. These systems cannot exactly chat like a human, but they often seem to. They can also retrieve and repackage information with a speed that humans never could. They can be thought of as digital assistants — like Siri or Alexa — that are better at understanding what you are looking for and giving it to you.

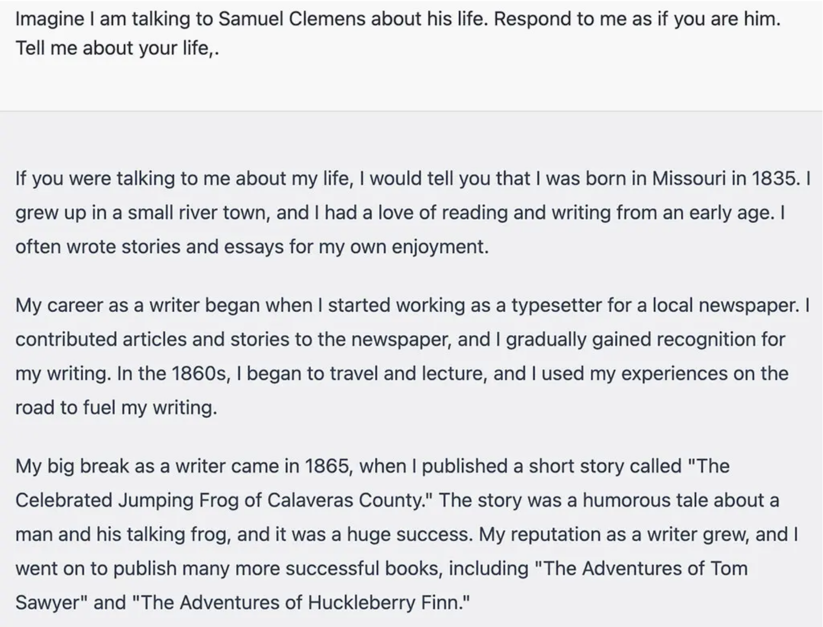


Jeremy Howard, a researcher in artificial intelligence, asked his young daughter to use a cutting-edge A.I. chatbot system. David Kelly for The New York Times

The new chatbots do this with what seems like complete confidence. But they do not always tell the truth. Sometimes, they even fail at simple arithmetic. They blend fact with fiction. And as they continue to improve, people could use them to generate and spread untruths.

Google recently built a system specifically for conversation, called LaMDA, or Language Model for Dialogue Applications. This spring, a Google engineer claimed it was sentient. It was not, but it captured the public's imagination.

Aaron Margolis, a data scientist in Arlington, Va., was among the limited number of people outside Google who were allowed to use LaMDA through an experimental Google app, AI Test Kitchen. He was consistently amazed by its talent for open-ended conversation. It kept him entertained. But he warned that it could be a bit of a fabulist — as was to be expected from a system trained from vast amounts of information posted to the internet.



Imagine I am talking to Samuel Clemens about his life. Respond to me as if you are him. Tell me about your life,.

If you were talking to me about my life, I would tell you that I was born in Missouri in 1835. I grew up in a small river town, and I had a love of reading and writing from an early age. I often wrote stories and essays for my own enjoyment.

My career as a writer began when I started working as a typesetter for a local newspaper. I contributed articles and stories to the newspaper, and I gradually gained recognition for my writing. In the 1860s, I began to travel and lecture, and I used my experiences on the road to fuel my writing.

My big break as a writer came in 1865, when I published a short story called "The Celebrated Jumping Frog of Calaveras County." The story was a humorous tale about a man and his talking frog, and it was a huge success. My reputation as a writer grew, and I went on to publish many more successful books, including "The Adventures of Tom Sawyer" and "The Adventures of Huckleberry Finn."

When Mr. Margolis prompted ChatGPT to chat as if it were Mark Twain (whose real name was Samuel Clemens), it erroneously stated that the title frog in a short story could talk.

"What it gives you is kind of like an Aaron Sorkin movie," he said. Mr. Sorkin wrote "The Social Network," a movie often criticized for stretching the truth about the origin of Facebook. "Parts of it will be true, and parts will not be true."

He recently asked both LaMDA and ChatGPT to chat with him as if it were Mark Twain. When he asked LaMDA, it soon described a meeting between Twain and Levi Strauss, and said the writer had worked for the bluejeans mogul while living in San Francisco in the mid-1800s. It seemed true. But it was not. Twain and Strauss lived in San Francisco at the same time, but they never worked together.

Scientists call that problem "hallucination." Much like a good storyteller, chatbots have a way of taking what they have learned and reshaping it into something new — with no regard for whether it is true.

LaMDA is what artificial intelligence researchers call a neural network, a mathematical system loosely modeled on the network of neurons in the brain. This is the same technology that translates between French and English on services like Google Translate and identifies pedestrians as self-driving cars navigate city streets.

A neural network learns skills by analyzing data. By pinpointing patterns in thousands of cat photos, for example, it can learn to recognize a cat.

Five years ago, researchers at Google and labs like OpenAI started designing neural networks that analyzed enormous amounts of digital text, including books, Wikipedia articles, news stories and online chat logs. Scientists call them "large language models." Identifying billions of distinct patterns in the way people connect words, numbers and symbols, these systems learned to generate text on their own.

Their ability to generate language surprised many researchers in the field, including many of the researchers who built them. The technology could mimic what people had written and combine disparate concepts. You could ask it to write a "Seinfeld" scene in which Jerry learns an esoteric mathematical technique called a bubble sort algorithm — and it would.

Andrej Karpathy
@karpathy · Follow

Best ChatGPT prompt so far 😂

Keith Wynroe @keithwynroe

I can't compete with this

12:21 AM · Dec 2, 2022

With ChatGPT, OpenAI has worked to refine the technology. It does not do free-flowing conversation as well as Google's LaMDA. It was designed to operate more like Siri, Alexa and other digital assistants. Like LaMDA, ChatGPT was trained on a sea of digital text culled from the internet.

As people tested the system, it asked them to rate its responses. Were they convincing? Were they useful? Were they truthful? Then, through a technique called reinforcement learning, it used the ratings to hone the system and more carefully define what it would and would not do.

"This allows us to get to the point where the model can interact with you and admit when it's wrong," said Mira Murati, OpenAI's chief technology officer. "It can reject something that is inappropriate, and it can challenge a question or a premise that is incorrect."

The method was not perfect. OpenAI warned those using ChatGPT that it "may occasionally generate incorrect information" and "produce harmful instructions or biased content." But the company plans to continue refining the technology, and reminds people using it that it is still a research project.

Google, Meta and other companies are also addressing accuracy issues. Meta recently removed an online preview of its chatbot, Galactica, because it repeatedly generated incorrect and biased information.

Experts have warned that companies do not control the fate of these technologies. Systems like ChatGPT, LaMDA and Galactica are based on ideas, research papers and computer code that have circulated freely for years.

Companies like Google and OpenAI can push the technology forward at a faster rate than others. But their latest technologies have been reproduced and widely distributed. They cannot prevent people from using these systems to spread misinformation.

Just as Mr. Howard hoped that his daughter would learn not to trust everything she read on the internet, he hoped society would learn the same lesson.

“You could program millions of these bots to appear like humans, having conversations designed to convince people of a particular point of view” he said. “I have warned about this for years. Now it is obvious that this is just waiting to happen.”

Cade Metz is a technology reporter and the author of “Genius Makers: The Mavericks Who Brought A.I. to Google, Facebook, and The World.” He covers artificial intelligence, driverless cars, robotics, virtual reality and other emerging areas. More about Cade Metz

A version of this article appears in print on , Section B, Page 1 of the New York edition with the headline: Chat Bots Can Amaze, But Also Lie