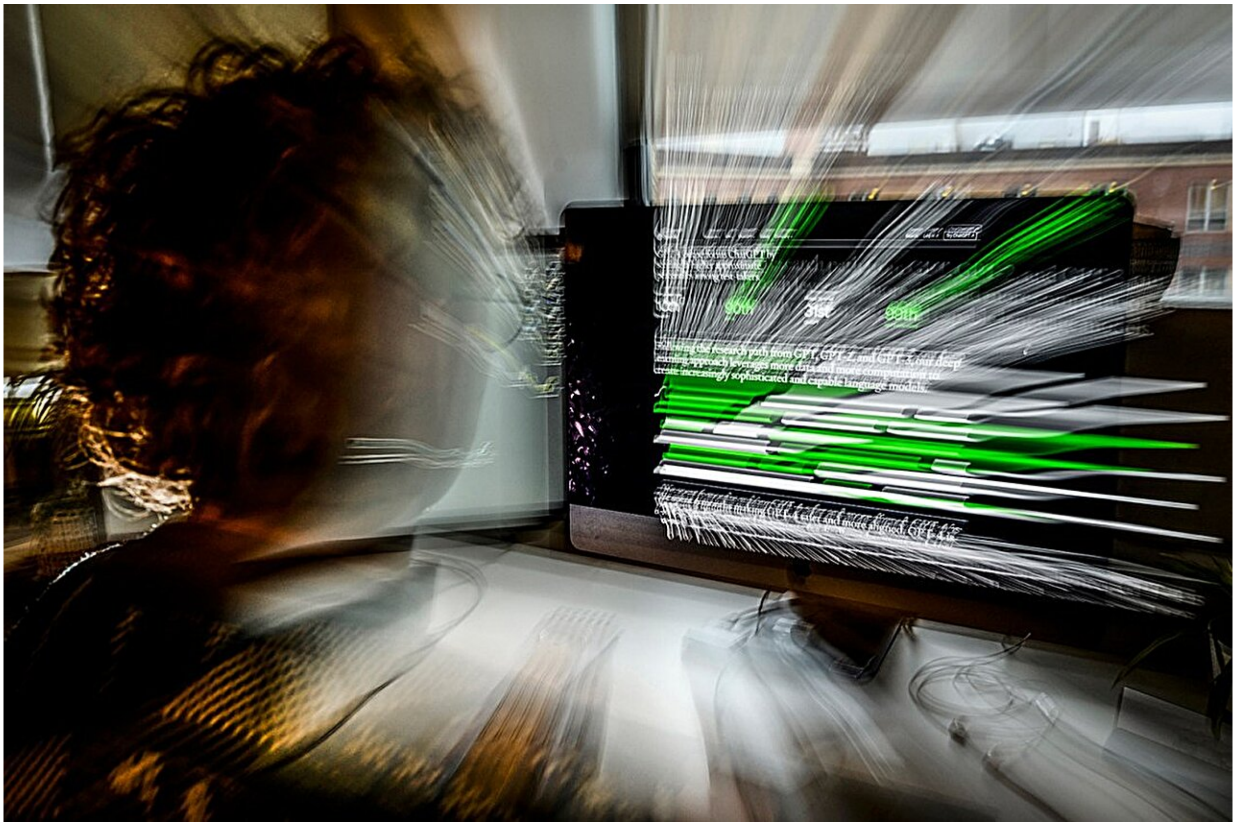
What are AI chatbots actually doing when they 'hallucinate?' Here's why experts don't like the term

November 13 2023, by Tanner Stening



A leading expert in the field is pushing back against the concept of “hallucination,” arguing that it gets much of how current AI models operate wrong. Credit: Matthew Modoono/Northeastern University

What are AI chatbots actually doing when they "hallucinate"? Does the term accurately capture why so-called generative AI tools—nearing ubiquity in many professional settings—sometimes generate false information when prompted?

As debate over the true nature, capacity and trajectory of AI applications simmers in the background, a leading expert in the field is pushing back against the concept of "hallucination," arguing that it gets much of how current AI models operate wrong.

"Generally speaking, we don't like the term because these models make errors—and we can explain why they make errors," says Usama Fayyad, executive director for the Institute for Experiential Artificial Intelligence at Northeastern University.

Fayyad says the term hallucination was popularized by Google in response to the launch of OpenAI's massively influential ChatGPT. While it serves as a compelling analog for the technology's human-like qualities and foibles, the term is something of a misnomer with potentially negative implications for the public's understanding of AI technology.

"When you say hallucinations, you're attributing too much to the model," Fayyad continues. "You're attributing intent; you're attributing consciousness; you're attributing a default mode of operating rationally; and you're attributing some form of understanding on the part of the machine."

Fayyad stresses that chatbots "don't have intent; [they] don't have understanding." He says the kinds of errors they make are not all that different from the errors inherent to any forecasting model—such as those used in economic or financial forecasts, where errors are readily anticipated and factored in appropriately.

Just how often chatbots "hallucinate" is still little-known, though some companies have been hard at work trying to quantify the error rates of the widely used large language models. One such company—a startup founded by former Google employees called Vectara—found that OpenAI's models hallucinated roughly 3% of the time, while a Google platform called "Palm chat" generated bogus info at a rate of 27%, according to the New York Times.

Further complicating matters is the fact that the auto-complete output produced by current generative AI models is highly dependent on the prompt, Fayyad says. Tweak the prompt even a tiny bit, and you get a very different result.

Byron Wallace, director for the data science program and the Sy and Laurie Sternberg Interdisciplinary Associate Professor in the Khoury College, once referred to these prompt designs as "not quite prompt engineering"—the process of designing inputs for chatbots—but more like "incantations and black magic."

Scaling back all of this hocus pocus, Fayyad wants to simplify the conversation around the potential application of generative AI tools.

"I could say—these models hallucinated; or, to be more precise, I could say, well, the model made an error, and we understand that these models make errors," Fayyad says.

To mix metaphors further, Fayyad makes the case that greater trust is needed between human beings and AI machines moving forward. He argues that "practitioners, users and organizations need to trust how a system reaches decisions, how it operates and the fact that it won't exhibit erratic, [unpredictable] or dangerous behavior."

"The topic of AI breeds mystery and ambiguity," he writes.

"Demystifying the technology and the behaviors exhibited by algorithms, good or bad, establishes real progress and creates valuable outcomes on all fronts: theoretical, academic, commercial and practical."

As it stands, large language models such as ChatGPT function as "glorified auto-complete" applications trained on huge amounts of digital text from online databases, articles and other sources. "They're just producing outputs just like any auto-complete device—your mobile phone or whatever."

"These models do not know the difference between a correct sequence and an error," Fayyad says. "Understanding where that error happens, and trying to recover from it—that is the very hard AI problem that we don't have very good solutions for today."

In an effort to rein in hallucinations, researchers have begun using other large language models to check the accuracy of various chatbots. Of course, those tools are capable of generating errors (hallucinations) as well, Fayyad notes.

He emphasizes the need for human beings to continue checking the output generated by these tools—a concept referred to as "human-in-the-loop."

"It leaves it to you—the user—to say, 'this auto-complete is not correct,'' and then fix it," he says.

Provided by Northeastern University

Citation: What are AI chatbots actually doing when they 'hallucinate?' Here's why experts don't like the term (2023, November 13) retrieved 2 May 2024 from https://techxplore.com/news/2023-11-ai-chatbots-hallucinate-experts-dont.html