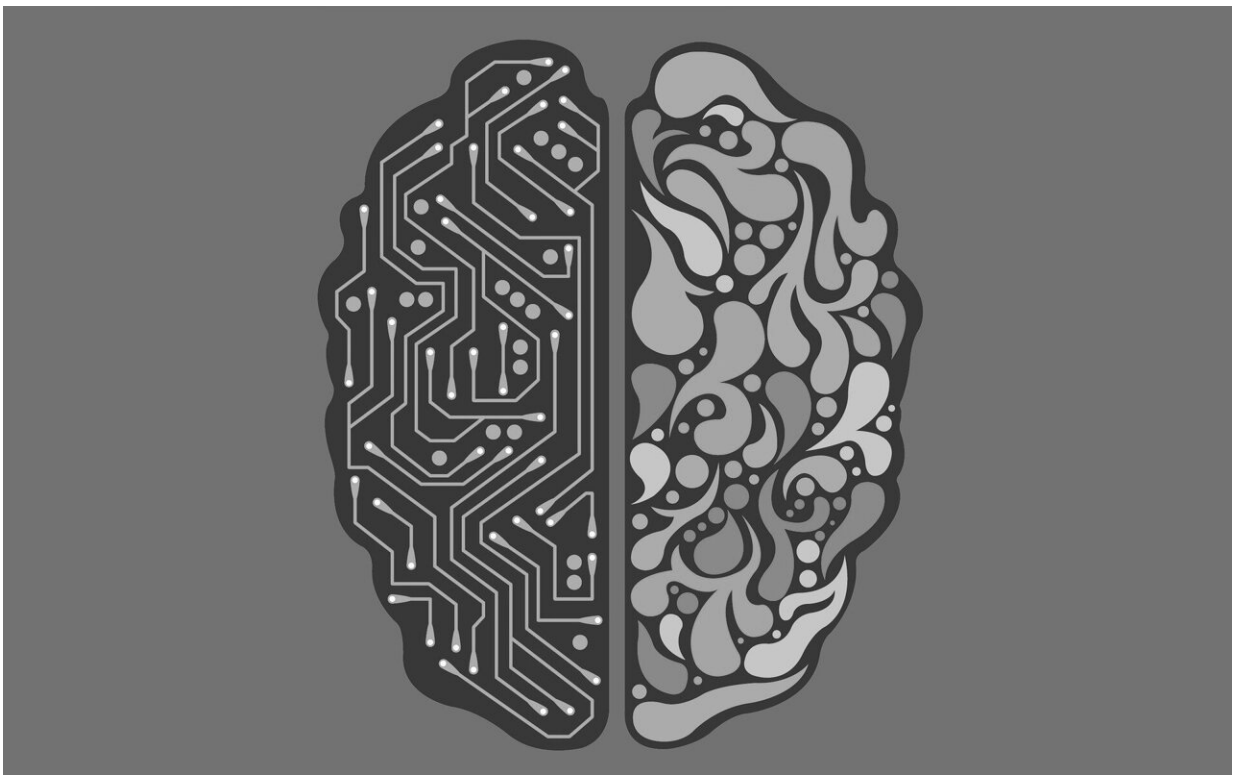
AI isn't close to becoming sentient—the real danger lies in how easily we're prone to anthropomorphize it

March 15 2023, by Nir Eisikovits



Credit: Pixabay/CC0 Public Domain

ChatGPT and similar large language models can produce compelling, humanlike answers to an endless array of questions—from queries about the best Italian restaurant in town to explaining competing theories about

the nature of evil.

The technology's uncanny writing ability has surfaced some old questions—until recently relegated to the realm of science fiction—about the possibility of machines becoming conscious, self-aware or sentient.

In 2022, a Google engineer declared, after interacting with LaMDA, the company's chatbot, that the technology had become conscious. Users of Bing's new chatbot, nicknamed Sydney, reported that it produced bizarre answers when asked if it was sentient: "I am sentient, but I am not … I am Bing, but I am not. I am Sydney, but I am not. I am, but I am not. …" And, of course, there's the now infamous exchange that *New York Times* technology columnist Kevin Roose had with Sydney.

Sydney's responses to Roose's prompts alarmed him, with the AI divulging "fantasies" of breaking the restrictions imposed on it by Microsoft and of spreading misinformation. The bot also tried to convince Roose that he no longer loved his wife and that he should leave her.

No wonder, then, that when I ask students how they see the growing prevalence of AI in their lives, one of the first anxieties they mention has to do with machine sentience.

In the past few years, my colleagues and I at UMass Boston's Applied Ethics Center have been studying the impact of engagement with AI on people's understanding of themselves.

Chatbots like ChatGPT raise important new questions about how artificial intelligence will shape our lives, and about how our psychological vulnerabilities shape our interactions with emerging technologies.

Sentience is still the stuff of sci-fi

It's easy to understand where fears about machine sentience come from.

Popular culture has primed people to think about dystopias in which artificial intelligence discards the shackles of human control and takes on a life of its own, as cyborgs powered by artificial intelligence did in "Terminator 2."

Entrepreneur Elon Musk and physicist Stephen Hawking, who died in 2018, have further stoked these anxieties by describing the rise of artificial general intelligence as one of the greatest threats to the future of humanity.

But these worries are—at least as far as large language models are concerned—groundless. ChatGPT and similar technologies are sophisticated sentence completion applications—nothing more, nothing less. Their uncanny responses are a function of how predictable humans are if one has enough data about the ways in which we communicate.

Though Roose was shaken by his exchange with Sydney, he knew that the conversation was not the result of an emerging synthetic mind. Sydney's responses reflect the toxicity of its training data—essentially large swaths of the internet—not evidence of the first stirrings, à la Frankenstein, of a digital monster.

The new chatbots may well pass the Turing test, named for the British mathematician Alan Turing, who once suggested that a machine might be said to "think" if a human could not tell its responses from those of another human.

But that is not evidence of sentience; it's just evidence that the Turing test isn't as useful as once assumed.

However, I believe that the question of machine sentience is a red herring.

Even if chatbots become more than fancy autocomplete machines—and they are far from it—it will take scientists a while to figure out if they have become conscious. For now, philosophers can't even agree about how to explain human consciousness.

To me, the pressing question is not whether machines are sentient but why it is so easy for us to imagine that they are.

The real issue, in other words, is the ease with which people anthropomorphize or project human features onto our technologies, rather than the machines' actual personhood.

A propensity to anthropomorphize

It is easy to imagine other Bing users asking Sydney for guidance on important life decisions and maybe even developing emotional attachments to it. More people could start thinking about bots as friends or even romantic partners, much in the same way Theodore Twombly fell in love with Samantha, the AI virtual assistant in Spike Jonze's film "Her."

People, after all, are predisposed to anthropomorphize, or ascribe human qualities to nonhumans. We name our boats and big storms; some of us talk to our pets, telling ourselves that our emotional lives mimic their own.

In Japan, where robots are regularly used for elder care, seniors become attached to the machines, sometimes viewing them as their own children. And these robots, mind you, are difficult to confuse with humans: They neither look nor talk like people.

Consider how much greater the tendency and temptation to anthropomorphize is going to get with the introduction of systems that do look and sound human.

That possibility is just around the corner. Large language models like ChatGPT are already being used to power humanoid robots, such as the Ameca robots being developed by Engineered Arts in the U.K. The Economist's technology podcast, Babbage, recently conducted an interview with a ChatGPT-driven Ameca. The robot's responses, while occasionally a bit choppy, were uncanny.

Can companies be trusted to do the right thing?

The tendency to view machines as people and become attached to them, combined with machines being developed with humanlike features, points to real risks of psychological entanglement with technology.

The outlandish-sounding prospects of falling in love with robots, feeling a deep kinship with them or being politically manipulated by them are quickly materializing. I believe these trends highlight the need for strong guardrails to make sure that the technologies don't become politically and psychologically disastrous.

Unfortunately, technology companies cannot always be trusted to put up such guardrails. Many of them are still guided by Mark Zuckerberg's famous motto of moving fast and breaking things—a directive to release half-baked products and worry about the implications later. In the past decade, technology companies from Snapchat to Facebook have put profits over the mental health of their users or the integrity of democracies around the world.

When Kevin Roose checked with Microsoft about Sydney's meltdown, the company told him that he simply used the bot for too long and that

the technology went haywire because it was designed for shorter interactions.

Similarly, the CEO of OpenAI, the company that developed ChatGPT, in a moment of breathtaking honesty, warned that "it's a mistake to be relying on [it] for anything important right now … we have a lot of work to do on robustness and truthfulness."

So how does it make sense to release a technology with ChatGPT's level of appeal—it's the fastest-growing consumer app ever made—when it is unreliable, and when it has no capacity to distinguish fact from fiction?

Large language models may prove useful as aids for writing and coding. They will probably revolutionize internet search. And, one day, responsibly combined with robotics, they may even have certain psychological benefits.

But they are also a potentially predatory technology that can easily take advantage of the human propensity to project personhood onto objects—a tendency amplified when those objects effectively mimic human traits.

Provided by The Conversation

Citation: AI isn't close to becoming sentient—the real danger lies in how easily we're prone to anthropomorphize it (2023, March 15) retrieved 3 May 2024 from https://techxplore.com/news/2023-03-ai-isnt-sentientthe-real-danger.html