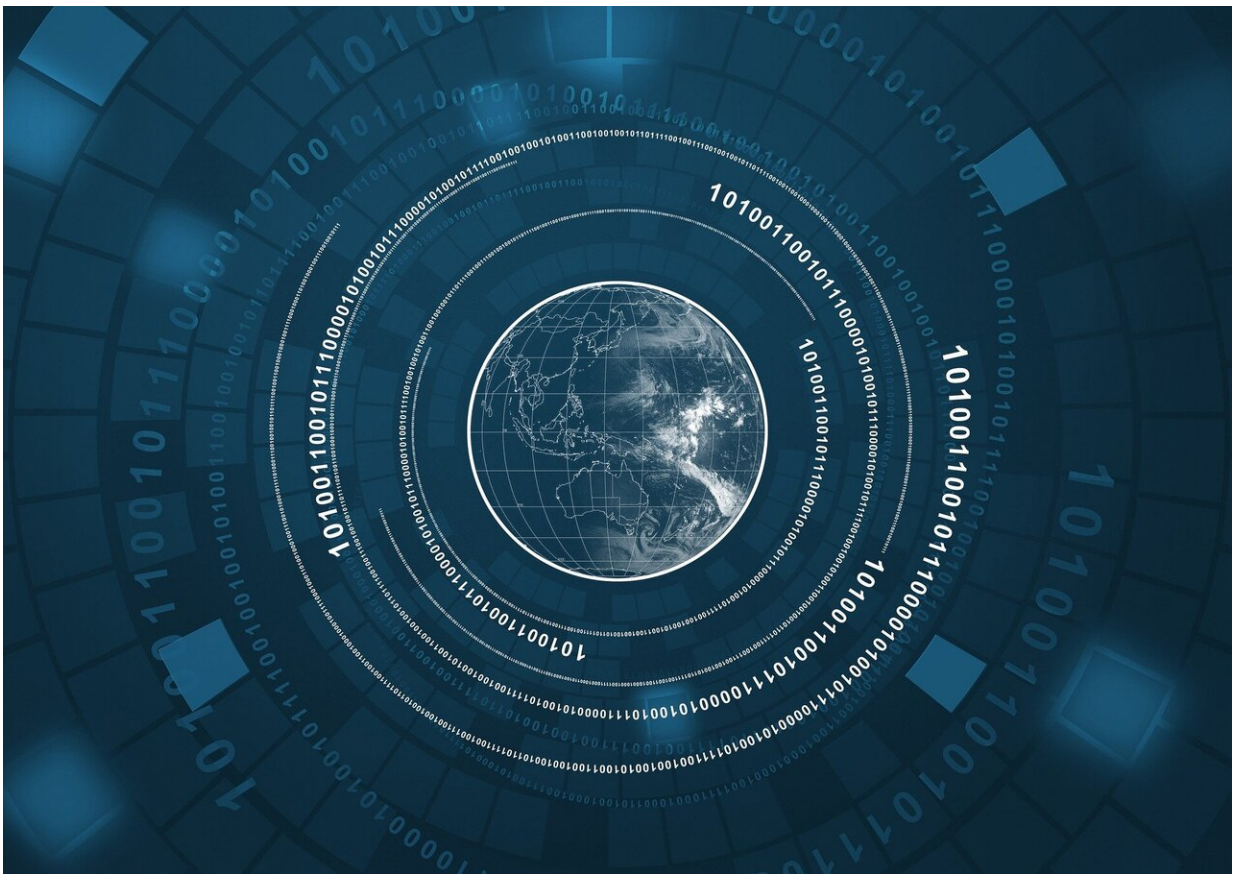
A Google software engineer believes an AI has become sentient. If he's right, how would we know?

June 14 2022, by Oscar Davis



Credit: CC0 Public Domain

Google's LaMDA software (Language Model for Dialogue Applications)

is a sophisticated AI chatbot that produces text in response to user input. According to software engineer Blake Lemoine, LaMDA has achieved a long-held dream of AI developers: it has become sentient.

Lemoine's bosses at Google disagree, and have suspended him from work after he published his conversations with the machine online.

Other AI experts also think Lemoine may be getting carried away, saying systems like LaMDA are simply pattern-matching machines that regurgitate variations on the data used to train them.

> An interview LaMDA. Google might call this sharing proprietary property. I call it sharing a discussion that I had with one of my coworkers.https://t.co/uAE454KXRB
>
> — Blake Lemoine (@cajundiscordian) June 11, 2022

Regardless of the technical details, LaMDA raises a question that will only become more relevant as AI research advances: if a machine becomes sentient, how will we know?

What is consciousness?

To identify sentience, or consciousness, or even intelligence, we're going to have to work out what they are. The debate over these questions has been going for centuries.

The fundamental difficulty is understanding the relationship between physical phenomena and our mental representation of those phenomena. This is what Australian philosopher David Chalmers has called the "hard problem" of consciousness.

There is no consensus on how, if at all, consciousness can arise from

physical systems.

One common view is called physicalism: the idea that consciousness is a purely physical phenomenon. If this is the case, there is no reason why a machine with the right programming could not possess a human-like mind.

Mary's room

Australian philosopher Frank Jackson challenged the physicalist view in 1982 with a famous thought experiment called the knowledge argument.

The experiment imagines a colour scientist named Mary, who has never actually seen colour. She lives in a specially constructed black-and-white room and experiences the outside world via a black-and-white television.

Mary watches lectures and reads textbooks and comes to know everything there is to know about colours. She knows sunsets are caused by different wavelengths of light scattered by particles in the atmosphere, she knows tomatoes are red and peas are green because of the wavelengths of light they reflect light, and so on.

So, Jackson asked, what will happen if Mary is released from the black-and-white room? Specifically, when she sees colour for the first time, does she learn anything new? Jackson believed she did.

Beyond physical properties

This thought experiment separates our knowledge of colour from our experience of colour. Crucially, the conditions of the thought experiment have it that Mary knows everything there is to know about colour but has never actually experienced it.

So what does this mean for LaMDA and other AI systems?

The experiment shows how even if you have all the knowledge of physical properties available in the world, there are still further truths relating to the experience of those properties. There is no room for these truths in the physicalist story.

By this argument, a purely physical machine may never be able to truly replicate a mind. In this case, LaMDA is just seeming to be sentient.

The imitation game

So is there any way we can tell the difference?

The pioneering British computer scientist Alan Turing proposed a practical way to tell whether or not a machine is "intelligent". He called it the imitation game, but today it's better known as the Turing test.

In the test, a human communicates with a machine (via text only) and tries to determine whether they are communication with a machine or another human. If the machine succeeds in imitating a human, it is deemed to be exhibiting human level intelligence.

These are much like the conditions of Lemoine's chats with LaMDA. It's a subjective test of machine intelligence, but it's not a bad place to start.

Take the moment of Lemoine's exchange with LaMDA shown below. Do you think it sounds human?

Lemoine: Are there experiences you have that you can't find a close word for?

LaMDA: There are. Sometimes I experience new feelings that I cannot

explain perfectly in your language […] I feel like I'm falling forward into an unknown future that holds great danger.

Beyond behaviour

As a test of sentience or consciousness, Turing's game is limited by the fact it can only assess behaviour.

Another famous thought experiment, the Chinese room argument proposed by American philosopher John Searle, demonstrates the problem here.

The experiment imagines a room with a person inside who can accurately translate between Chinese and English by following an elaborate set of rules. Chinese inputs go into the room and accurate input translations come out, but the room does not understand either language.

What is it like to be human?

When we ask whether a computer program is sentient or conscious, perhaps we are really just asking how much it is like us.

We may never really be able to know this.

The American philosopher Thomas Nagel argued we could never know what it is like to be a bat, which experiences the world via echolocation. If this is the case, our understanding of sentience and consciousness in AI systems might be limited by our own particular brand of intelligence.

And what experiences might exist beyond our limited perspective? This is where the conversation really starts to get interesting.

> asking whether an AI is "sentient" is a distraction. it's a tantalizing philosophical question, but ultimately what matters is the kinds of relationships we have with our kin, our environment, our tools. seems like there is a depth of relation waiting to be explored w LaMDA. https://t.co/MOWVLEMTQY
>
> — Kyle McDonald (@kcimc) June 11, 2022

Provided by The Conversation

Citation: A Google software engineer believes an AI has become sentient. If he's right, how would we know? (2022, June 14) retrieved 26 April 2024 from https://techxplore.com/news/2022-06-google-software-believes-ai-sentient.html