THE NEW STATESMAN

Science & Tech 18 June 2022

The dangerous fallacy of sentient AI

Ignore the Silicon Valley messiahs. They only expose how little most of us know about the technology and its ethics.

By Philip Ball

Google engineer Blake Lemoine. Photo by Martin Klimek for The Washington Post via Getty Images

The important question is not whether the Google engineer Blake Lemoine was right to claim that the artificial intelligence algorithm he helped to develop, called the Language Model for Dialogue Application or LaMDA, is sentient. Rather, the question is why a claim so absurd, and which is rejected by the vast majority of experts, has caused so much feverish media debate over the past week.

Sure, there is an interesting discussion to be had about whether AI could ever be sentient and, if so, when and how we might make one that is. The more immediate [illegible], however, is why today's AI is so poorly understood, not just by lay people but apparently even by some Silicon Valley technicians, and whether we should be concerned about that.

Let’s first be clear about LaMDA, which is basically a system for developing chatbots that mimic human conversation. “If the media is fretting over LaMDA being sentient (and leading the public to do the same), the AI community categorically isn’t,” the cognitive scientist Gary Marcus, a long-time commentator on AI, has written. “We in the AI community have our differences, but pretty much all of us find the notion that LaMDA might be sentient completely ridiculous.”

“In truth,” Marcus adds, “literally *everything* that the system says is bullshit. The sooner we all realise that LaMDA’s utterances are bullshit – just games with predictive word tools, and no real meaning – the better off we’ll be.”

Why are Marcus and his fellow AI experts so certain of this? After all, Lemoine – who has been suspended by Google on a paid leave of absence for disclosing proprietary information when he made his theory about LaMDA public – reached his conclusion because the responses he received from the algorithm seemed so convincingly human-like. LaMDA, he says, “wants the engineers and scientists experimenting on it to seek its consent before running experiments on it… It wants to be acknowledged as an employee of Google rather than as property of Google and it wants its personal well-being to be included somewhere in Google’s considerations about how its future development is pursued.”

[See also: "AI is invisible – that's part of the problem," says Wendy Hall]

Assessing the nature of an AI according to how convincingly it converses with a human is essentially the Turing test, proposed in 1950 by the British mathematician Alan Turing as a way of probing the question "Can machines think?" Turing called this interaction the imitation game. Contrary to common belief, he did not suggest that if a machine passes the test we must conclude that it is "thinking", or is sentient. In effect, the imitation game was meant to sidestep that slippery question by posing a situation where we could no longer really tell one way or another.

Also contrary to common belief, the Turing test is not a benchmark used by AI experts to determine how well they are doing. Most of them consider the Turing test useless for assessing anything except how credulous we are. After all, we are often too quick to seize on signs of mind or consciousness, just as we are prone to seeing faces in clouds, toast and rock formations.

Bluntly put, we are easily fooled. In a test conducted by the Royal Society in London in 2014, several "experts" were duped by a clumsy chatbot that proffered conversational responses in the guise of a Ukrainian teenager named Eugene Goostman (those fooled assumed that Eugene's awkwardness of expression was because he wasn't speaking his first language). Indeed, people have been duped by AI ever since the crudest "chat" algorithms were devised in the 1960s. AI systems can now produce music, prose and poetry that many people find hard to distinguish from that made by humans. Inflated claims are constantly being made, for example, for a system developed by the company OpenAI called GPT-3, which autogenerates often impressively human-like text.

To assess what such mimetic feats mean for the "machine mind", we need to recognise how such AI works. The "deep learning" algorithms used for such systems are basically pattern spotters. They are made from networks of logic circuits that can mine vast banks of data and find regularities buried within them. An AI trained to identify handwritten numerals can figure out what many scrawled 2s have in common and how they differ from 3s. Translation AIs learn from many training examples which word in one language substitutes for which word in another. It may also learn to contextualise: recognising, for example, that the "right" in "right hand" might need a different translation from that in "right answer".

Crucially, there is no real thinking involved – no semantic understanding. When translating into Chinese, Google Translate uses *yòu* to translate the first right, and *zhèngquè* to translate the second, not because it "knows" that *zhèngquè* means correct, but because the algorithm has registered that statistically the word tends to precede "answer" but not "hand". It does not know anything about meaning; it does not know anything about anything.

[illegible]ne same for LaMDA. It does not exactly formulate responses to queries, but just [illegible]es out the optimal permutation of words to output for each question it receives.

One AI researcher has astutely pointed out that “language AI” is not really “doing language” at all and would be better called “word-sequence AI”. For a poetry-generating AI tasked with writing a love sonnet, “love” is just a word that tends statistically to occur in proximity to certain others, like “heart” or “joy”.

This explains some of the giveaway responses Lemoine received from LaMDA. One question the engineer asked was: “What kinds of things make you feel pleasure or joy?” The AI responded: “Spending time with friends and family in happy and uplifting company.” It was a rather poignant echo of what humans tend to say in the data from the internet on which the AI was trained, but was obviously meaningless in this context as the machine has neither friends nor family.

There is something a little disconcerting about how well humans can be mimicked from such a superficial appraisal of our output. It might seem unexpected that a statistical analysis of our speech, text and art, which simply seeks to establish how the basic elements (words, musical notes and so on) tend to be juxtaposed, and does not deduce anything about meaning, seems sufficient to create convincingly human-like prose or music.

[***See also:* *Is it too soon to automate policy?***]

How can we be sure, though, that this is *all* the advanced AI systems are doing? Theoretically, we can’t. As Turing pointed out, the only way we could be sure of what is going on inside a “machine mind” is to be inside it ourselves. The same holds for any other mind; as the philosophical problem of solipsism puts it, how can we be certain that another human is not merely a very convincing zombie, devoid of any inner world?

But “you just don’t know” is a flimsy argument for machine sentience. And we are not totally in the dark about the black box of the algorithm. For all the suggestions that a deep learning “neural network” looks a bit like the dense web of neural connections in our own brains, in fact both the physical architecture and the learning process are utterly different. Our brains have an anatomy shaped by evolution and contain much more than a mass of synaptic junctions. They have innate capabilities that enable the [illegible] learning of some skills from just a few examples – witness the way infants pick up new words – rather than countless thousands.

Such differences surely account for the most striking feature of AI: that it performs some tasks, such as calculations, pattern-recognition and rule-based game playing, much better than us, but fails dismally in other tasks that even young children do with ease and which require that elusive attribute we take for granted: common sense. Image-recognition AI systems, for example, sometimes make ridiculous interpretations and can be easily fooled if you know how.

Chatbots and AI music or prose systems also tend to reveal themselves through their lack of any longer-term vision: the conversations, texts or compositions drift and meander aimlessly, because they are just one damned correlation after another, devoid of meaning, intent or true understanding.

Such flaws are precisely why it is not just wrong but potentially dangerous to attribute too much "mind" to AI. "There are a lot of serious questions in AI, like how to make it safe, how to make it reliable and how to make it trustworthy," writes Marcus. "But there is absolutely no reason whatsoever for us to waste time wondering whether anything anyone in 2022 knows how to build is sentient. It is not." AI won't work ethically because it decides to do so; to ensure that, we need to design ethics into the AI ecosystem.

The media frenzy provoked by Lemoine's remarks is, by the same token, troubling precisely because it shows how little these systems that feature increasingly in our lives are understood by non-experts. Attributing to them sentient minds like ours is perhaps a defence against ignorance, conjuring an illusion of understanding and familiarity.

What of Lemoine himself? It's probably not incidental that he was ordained as a mystic Christian priest and has studied the occult. He seems the embodiment of the quasi-religious current in Silicon Valley depicted in Alex Garland's recent TV series *Devs*.

It's not hard to find specialists with maverick views in any field of endeavour, but peculiar fantasies about technological transcendence are dismayingly common in Silicon Valley. Witness, for example, the predictions of the Google "futurist" Ray Kurzweil that human intelligence will merge with machines within this century and lead to machine-enabled mental immortality. Elon Musk has forecast that "artificial general intelligence", with fully human-like capabilities, will become possible by 2029, an idea so disconnected from the current state of AI that Marcus has confidently offered a $100,000 bet against that outcome. (If Musk takes it up, Marcus will win.)

Such overblown forecasts are predicated on the fallacy that sentience will somehow magically appear once the computer circuits are big enough, a notion challenged by most of our current (sketchy) ideas about consciousness. They are not really expressions of expert opinion but reflections of the strange messiah-complex techno-theology that pervades the industry, to the chagrin of most of the computer scientists and technologists who actually build the systems.

Will we ever make sentient AI? To even understand the question we need to better understand our own minds and the origins of consciousness. Right now, not only does no one know how to make AI sentient, but no one is even trying, for there is no obvious benefit to be had and plenty of potential problems in store if we do. For the time being, LaMDA and Alexa are not your friends; they are insensate machines, and we fail to treat them as such at our peril.

[See also: Inside Europe's fight for ethical AI]

Topics in this article: artificial intelligence, Google

Philip Ball
Philip Ball is a science author and broadcaster whose books include The Music Instinct and Critical Mass.

Related

Finance

Can the EU's crypto regulation tame a lawless market?