Subscribe

The Atlantic

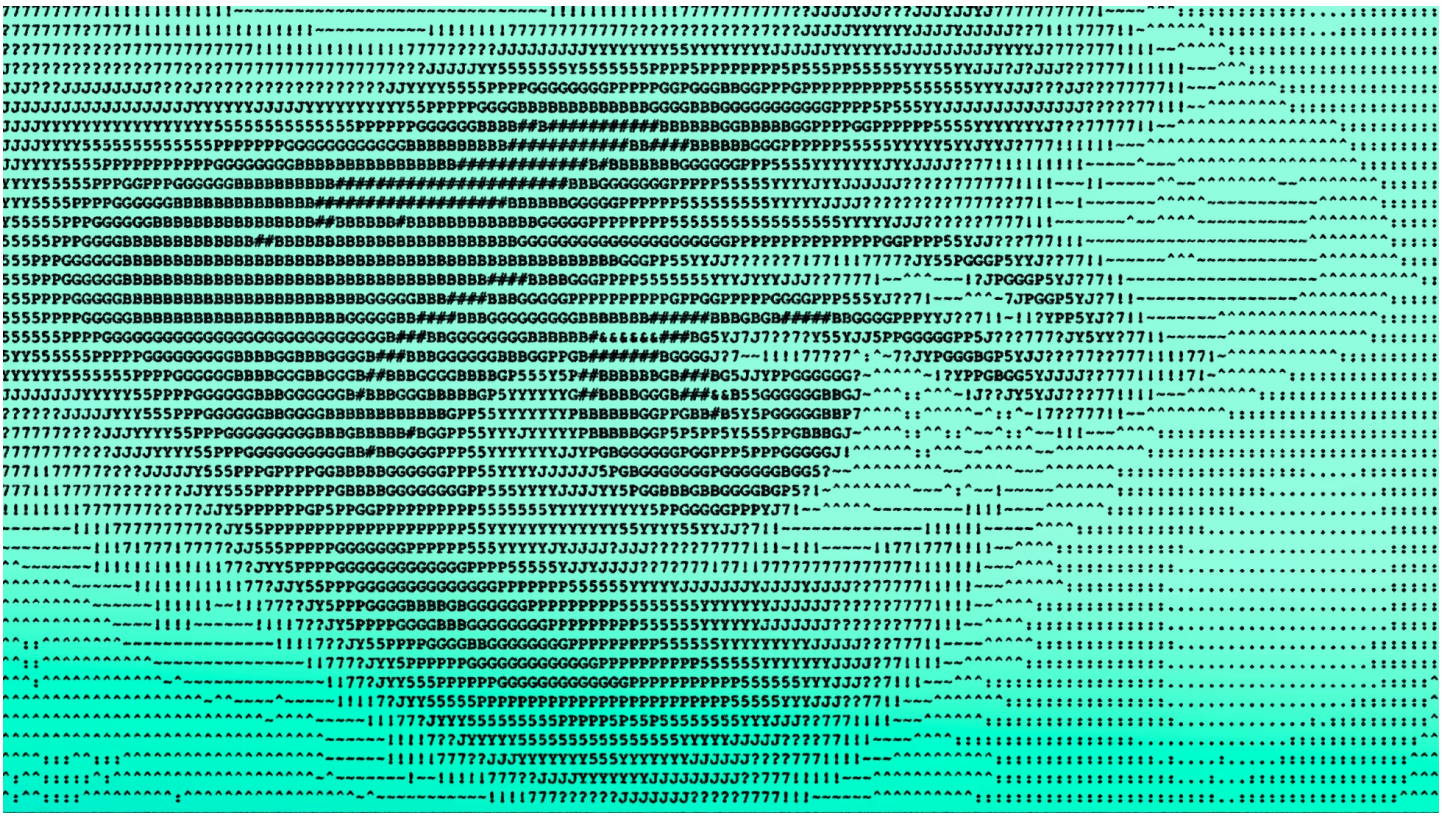
The Supreme Court has overturned *Roe v. Wade*. Read *The Atlantic*'s full coverage of the abortion debate.

IDEAS

How a Google Employee Fell for the Eliza Effect

Determining who and what is or is not sentient is one of the defining questions of almost any moral code.

By Brian Christian



The Atlantic; Getty

JUNE 21, 2022

SHARE

About the author: *Brian Christian is the author of three books about the human and philosophical implications of computer science:* The Most Human Human*,* Algorithms to Live By*, and* The Alignment Problem.

A Google employee named Blake Lemoine was put on leave recently after claiming that one of Google's artificial-intelligence language models, called LaMDA (Language Models for Dialogue Applications), is sentient. He went public with his concerns, sharing his text conversations with LaMDA. At one point, Lemoine asks, "What does the word 'soul' mean to you?" LaMDA answers, "To me, the soul is a concept of the animating force behind consciousness and life itself."

"I was inclined to give it the benefit of the doubt," Lemoine explained, citing his religious beliefs. "Who am I to tell God where he can and can't put souls?"

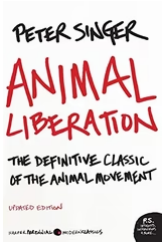
I do not believe that Lemoine's text exchanges are evidence of sentience. Behind the question of what these transcripts do or do not prove, however, is something much deeper and more profound: an invitation to revisit the humbling, fertile, and in-flux question of sentience itself.

ENJOY A YEAR OF UNLIMITED ACCESS TO THE ATLANTIC—INCLUDING EVERY STORY ON OUR SITE AND APP, SUBSCRIBER NEWSLETTERS, AND MORE.

Become a Subscriber

Read: Google's 'sentient' chatbot is our self-deceiving future

As the language-model catchphrase goes, let's think step-by-step.



Animal Liberation: The Definitive Classic Of The Animal Movement

PETER SINGER,
HARPER PERENNIAL MODERN CLASSICS

BUY BOOK

When you buy a book using a link on this page, we receive a commission. Thank you for supporting *The Atlantic*.

The first chatbot—a program designed to mimic human conversation—was called Eliza, written by the MIT professor Joseph Weizenbaum in the 1960s. As the story goes, his secretary came to believe that she was having meaningful dialogues with the system, despite the program's incredibly simple logic (mostly reflecting a user's statements back in the form of a question), and despite Weizenbaum's insistence that there was truly nothing more to it than that. This form of anthropomorphism has come to be known as the Eliza effect.

Lemoine—who seems, as far as I can tell, like a very thoughtful and kindhearted person of sincere convictions—was, I believe, a victim of the Eliza effect. LaMDA, like many other "large language models" (LLMs) of today, is a kind of autocomplete on steroids. It has been trained to fill in the blanks of missing words within an enormous linguistic corpus, then it is "fine-tuned" with further training specific to text dialogue. What these systems can do is breathtaking and sublime. I am more inclined than many to view LLMs' uncanny facility with language as evidence of some form of at least partially "real" (as opposed to "fake") linguistic understanding, for instance.

However, when LaMDA is asked by Lemoine to describe its "soul," it is not speaking "for itself"; it is autocompleting his prompt just as it would fill in the blanks of a science-fiction screenplay, say, or a Dadaist limerick, or a tech-support manual in the style of Chaucer.

DON'T MISS WHAT MATTERS. SIGN UP FOR THE ATLANTIC DAILY NEWSLETTER.

Enter your email Sign Up

What may sound like introspection is just the system improvising in an introspective verbal style, "Yes, and"–ing Lemoine's own thoughtful questions.

LaMDA fooled Lemoine. Does it follow that LaMDA "passes the Turing test" in a more general sense? That is, does LaMDA exhibit sufficiently human-seeming conversation that people consistently fail to distinguish it from the real thing?

RECOMMENDED READING



What I Learned When I Rented My Parents' Former Home as an Airbnb

THAD RUSSELL



Human History Gets a Rewrite

WILLIAM DERESIEWICZ



The Key to Escaping the Couple-Envy Trap

JOSHUA COLEMAN

Google could find out. It could hire, say, 30 crowdworkers to act as judges and 30 to act as human control subjects, and just have at it. Each judge would have one conversation with a human, one with LaMDA, and would then have to decide which was which. We'd have the results in 15 minutes. Following Alan Turing's 1950 paper, anything less than 70 percent accuracy by the judges would constitute the machines

"passing," so LaMDA would need to fool just nine of the 30 judges to pass the Turing test. If I had to, I'd bet (though not a lot) that LaMDA would, indeed, fool nine or more of the judges. Perhaps you disagree. But there's no need to argue, because finding out would be trivially easy.

Read: Artificial intelligence is misreading human emotion

Turing proposed his test—originally called the Imitation Game—as an empirical substitute for the more theoretical question of "Can machines think?" As Turing foresaw, language, and particularly conversation, has proved indeed to be a versatile medium for probing a diverse array of behaviors and capabilities. Conversation is still useful for testing the limits of today's LLMs. But as machines seem clearly to be succeeding ever more adeptly at the Imitation Game, the question of sentience, the true crux of the issue, begins to stand more apart from mere verbal facility.

MAKE YOUR INBOX MORE INTERESTING WITH NEWSLETTERS FROM YOUR FAVORITE ATLANTIC WRITERS.

Browse Newsletters

Sentience—of humans, of babies, of fetuses, of animals, of plants, of machines—has been debated for millennia. We have, in fact, learned a considerable amount about the neuroscience of consciousness, much of it unintuitive and surprising, just in the past several decades. Our collective understanding of these things has shifted considerably even within my lifetime.

In the 1940s, studies showing that newborn babies do not retract their limbs from pinpricks suggested that they did not feel pain, and this shifted medical consensus away from anesthetizing infants during surgery. In the late 1980s, further evidence—

of their stress hormones as well as brain development—overturned this view, making clear that anesthesia was ethically necessary.

In 1981, a presidential commission under Ronald Reagan held meetings including philosophers, theologians, and neuroscientists, who debated "whole brain" versus "higher brain" theories of death. Their report became the foundation for ending cardiopulmonary definitions of death in medical and legal settings, and it shaped the system of organ donation that we have today. The exact criteria for brain death have evolved in significant ways from the 1960s to the present, and many countries differ considerably.

This is very much a sprawling, open frontier. We are still learning about the differences between locked-in syndrome and persistent vegetative state; we are still learning about split-brain syndrome and blindsight and the extent to which we are conscious while dreaming. We are still learning about how early in utero a fetus develops the ability to feel, and how early in its lifetime a baby learns to form memories and recall past experiences.

It's also strange to me that people seem to have such strong views about nonhuman sentience when philosophers have been arguing for millennia about whether animals are sentient, and animal rights are nowhere near a settled ethical issue today.

From the July/August 2008 issue: Is Google making us stupid?

Descartes, in the 1630s, cut living animals open for research without compunction, and wrote that "there is none that leads weak minds further from the straight path of virtue than that of imagining that the souls of beasts are of the same nature as our own." Skipping ahead to the late 20th century, the influential 1975 book *Animal Liberation*, by the philosopher Peter Singer, argued for a totally different conception of the experiences and rights of animals. His work has helped spur concern for animal welfare among academic philosophers and the public at large. Most contemporary ethical philosophers I know regard factory farming, for instance, as one of the great moral travesties of our time, if not the greatest. However, the debate continues: Some philosophers argue that animals are conscious, some that they are not conscious, and

some that animals do (or, in some cases, do not) deserve moral consideration regardless of whether they are conscious.

I believe that AI systems can, in principle, be "conscious"/"sentient"/"self-aware"/moral agents/moral patients—if only because I have not seen any compelling arguments that they cannot. Those arguments would require an understanding of the nature of our *own* consciousness that we simply don't have.

The goalposts for what AI "can't do" are moving, these days, at a stunning rate. Progress in understanding the neuroscience of consciousness is moving at a comparatively glacial pace, but the revelations are no less stunning. And the Overton window on things such as nonhuman sentience—informed by both—is perceptibly shifting, as perhaps it should.

Determining who and what is or is not sentient is one of the defining questions of almost any moral code. And yet, despite its utterly central position in our ethics, this question is utterly mysterious. We understand it better than we did a generation ago, which is thrilling. Let this episode be, then, an invitation to go to the literature, to learn something surprising or uncomfortable, and to reckon with how little we understand about the ultimate mystery at the center of both the physical and moral universe.

Animal Liberation: The Definitive Classic Of The Animal Movement

PETER SINGER, HARPER PERENNIAL MODERN CLASSICS

BUY BOOK ⌄
