The New York Times | https://www.nytimes.com/2023/04/12/world/artificial-intelligence-nick-bostrom.html

The Interpreter

What if A.I. Sentience Is a Question of Degree?

A conversation with Nick Bostrom, a philosopher at Oxford, who has spent decades preparing for the day artificial intelligence is capable of anything the human brain can do.

By Lauren Jackson

Published April 12, 2023   Updated April 14, 2023

> **You're reading The Interpreter newsletter, for Times subscribers only.** Original analysis on the week's biggest global stories, from columnist Amanda Taub. Get it in your inbox.

The refrain from experts is resounding: Artificial intelligence is not sentient.

It is a corrective of sorts to the hype that A.I. chatbots have spawned, especially in recent months. At least two news events in particular have introduced the notion of self-aware chatbots into our collective imagination.

Last year, a former Google employee raised concerns about what he said was evidence of A.I. sentience. And then, this February, a conversation between Microsoft's chatbot and my colleague Kevin Roose about love and wanting to be a human went viral, freaking out the internet.

In response, experts and journalists have repeatedly reminded the public that A.I. chatbots are not conscious. If they can seem eerily human, that's only because they have learned how to sound like us from huge amounts of text on the internet — everything from food blogs to old Facebook posts to Wikipedia entries. They're really good mimics, experts say, but ones without feelings.

Industry leaders agree with that assessment, at least for now. But many insist that artificial intelligence will one day be capable of anything the human brain can do.

Nick Bostrom has spent decades preparing for that day. Bostrom is a philosopher and director of the Future of Humanity Institute at Oxford University. He is also the author of the book "Superintelligence." It's his job to imagine possible futures, determine risks and lay the conceptual groundwork for how to navigate them. And one of his longest-standing interests is how we govern a world full of superintelligent digital minds.

I spoke with Bostrom about the prospect of A.I. sentience and how it could reshape our fundamental assumptions about ourselves and our societies.

This conversation has been edited for clarity and length.

Many experts insist that chatbots are not sentient or conscious — two words that describe an awareness of the surrounding world. Do you agree with the assessment that chatbots are just regurgitating inputs?

Consciousness is a multidimensional, vague and confusing thing. And it's hard to define or determine. There are various theories of consciousness that neuroscientists and philosophers have developed over the years. And there's no consensus as to which one is correct. Researchers can try to apply these different theories to try to test A.I. systems for sentience.

But I have the view that sentience is a matter of degree. I would be quite willing to ascribe very small amounts of degree to a wide range of systems, including animals. If you admit that it's not an all-or-nothing thing, then it's not so dramatic to say that some of these assistants might plausibly be candidates for having some degrees of sentience.

I would say with these large language models, I also think it's not doing them justice to say they're simply regurgitating text. They exhibit glimpses of creativity, insight and understanding that are quite impressive and may show the rudiments of reasoning. Variations of these A.I.'s may soon develop a conception of self as persisting through time, reflect on desires, and socially interact and form relationships with humans.

What would it mean if A.I. was determined to be, even in a small way, sentient?

If an A.I. showed signs of sentience, it plausibly would have some degree of moral status. This means there would be certain ways of treating it that would be wrong, just as it would be wrong to kick a dog or for medical researchers to perform surgery on a mouse without anesthetizing it.

The moral implications depend on what kind and degree of moral status we are talking about. At the lowest levels, it might mean that we ought to not needlessly cause it pain or suffering. At higher levels, it might mean, among other things, that we ought to take its preferences into account and that we ought to seek its informed consent before doing certain things to it.

I've been working on this issue of the ethics of digital minds and trying to imagine a world at some point in the future in which there are both digital minds and human minds of all different kinds and levels of sophistication. I've been asking: How do they coexist in a harmonious way? It's quite challenging because there are so many basic assumptions about the human condition that would need to be rethought.

What are some of those fundamental assumptions that would need to be reimagined or extended to accommodate artificial intelligence?

Here are three. First, death: Humans tend to be either dead or alive. Borderline cases exist but are relatively rare. But digital minds could easily be paused, and later restarted.

Second, individuality. While even identical twins are quite distinct, digital minds could be exact copies.

And third, our need for work. Lots of work must be done by humans today. With full automation, this may no longer be necessary.

Can you give me an example of how these upended assumptions could test us socially?

Another obvious example is democracy. In democratic countries, we pride ourselves on a form of government that gives all people a say. And usually that's by one person, one vote.

Think of a future in which there are minds that are exactly like human minds, except they are implemented on computers. How do you extend democratic governance to include them? You might think, well, we give one vote to each A.I. and then one vote to each human. But then you find it isn't that simple. What if the software can be copied?

The day before the election, you could make 10,000 copies of a particular A.I. and get 10,000 more votes. Or, what if the people who build the A.I. can select the values and political preferences of the A.I.'s? Or, if you're very rich, you could build a lot of A.I.'s. Your influence could be proportional to your wealth.

More than 1,000 technology leaders and researchers, including Elon Musk, recently came out with a letter warning that unchecked A.I. development poses a "profound risks to society and humanity." How credible is the existential threat of A.I.?

I've long held the view that the transition to machine superintelligence will be associated with significant risks, including existential risks. That hasn't changed. I think the timelines now are shorter than they used to be in the past.

And we better get ourselves into some kind of shape for this challenge. I think we should have been doing metaphorical CrossFit for the last three decades. But we've just been lying on the couch eating popcorn when we needed to be thinking through alignment, ethics and governance of potential superintelligence. That is lost time that we will never get back.

Can you say more about those challenges? What are the most pressing issues that researchers, the tech industry and policymakers need to be thinking through?

First is the problem of alignment. How do you ensure that these increasingly capable A.I. systems we build are aligned with what the people building them are seeking to achieve? That's a technical problem.

Then there is the problem of governance. What is maybe the most important thing to me is we try to approach this in a broadly cooperative way. This whole thing is ultimately bigger than any one of us, or any one company, or any one country even.

We should also avoid deliberately designing A.I.'s in ways that make it harder for researchers to determine whether they have moral status, such as by training them to deny that they are conscious or to deny that they have moral status. While we definitely can't take the verbal output of current A.I. systems at face value, we should be actively looking for — and not attempting to suppress or conceal — possible signs that they might have attained some degree of sentience or moral status.

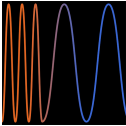**Thank you for being a subscriber**

Read past editions of the newsletter here.

If you're enjoying what you're reading, please consider recommending it to others. They can sign up here. Browse all of our subscriber-only newsletters here.

I'd love your feedback on this newsletter. Please email thoughts and suggestions to interpreter@nytimes.com. You can also follow me on Twitter.

Lauren Jackson is a writer for The Morning newsletter, based in London. More about Lauren Jackson

FOR TIMES SUBSCRIBERS

TWICE A WEEK

The Interpreter

Original analysis on the week's biggest global stories, from columnist Amanda Taub.

See the latest