MIND

How to Tell if Your A.I. Is Conscious

In a new report, scientists offer a list of measurable qualities that might indicate the presence of some presence in a machine.

By Oliver Whang

Sept. 18, 2023

Have you ever talked to someone who is "into consciousness?" How did that conversation go? Did they make a vague gesture in the air with both hands? Did they reference the Tao Te Ching or Jean-Paul Sartre? Did they say that, actually, there's nothing scientists can be sure about, and that reality is only as real as we make it out to be?

The fuzziness of consciousness, its imprecision, has made its study anathema in the natural sciences. At least until recently, the project was largely left to philosophers, who often were only marginally better than others at clarifying their object of study. Hod Lipson, a roboticist at Columbia University, said that some people in his field referred to consciousness as "the C-word." Grace Lindsay, a neuroscientist at New York University, said, "There was this idea that you can't study consciousness until you have tenure."

Nonetheless, a few weeks ago, a group of philosophers, neuroscientists and computer scientists, Dr. Lindsay among them, proposed a rubric with which to determine whether an A.I. system like ChatGPT could be considered conscious. The report, which surveys what Dr. Lindsay calls the "brand-new" science of consciousness, pulls together elements from a half-dozen nascent empirical theories and proposes a list of measurable qualities that might suggest the presence of some presence in a machine.

For instance, recurrent processing theory focuses on the differences between conscious perception (for example, actively studying an apple in front of you) and unconscious perception (such as your sense of an apple flying toward your face). Neuroscientists have argued that we unconsciously perceive things when electrical signals are passed from the nerves in our eyes to the primary visual cortex and then to deeper parts of the brain, like a baton being handed off from one cluster of nerves to another. These perceptions seem to become conscious when the baton is passed back, from the deeper parts of the brain to the primary visual cortex, creating a loop of activity.

Another theory describes specialized sections of the brain that are used for particular tasks — the part of your brain that can balance your top-heavy body on a pogo stick is different from the part of your brain that can take in an expansive landscape. We're able to put all this information together (you can bounce on a pogo stick while appreciating a nice view), but only to a certain extent (doing so is difficult). So neuroscientists have postulated the existence of a "global workspace" that allows for control and coordination over what we pay attention to, what we remember, even what we perceive. Our consciousness may arise from this integrated, shifting workspace.

But it could also arise from the ability to be aware of your own awareness, to create virtual models of the world, to predict future experiences and to locate your body in space. The report argues that any one of these features could, potentially, be an essential part of what it means to be conscious. And, if we're able to discern these traits in a machine, then we might be able to consider the machine conscious.

One of the difficulties of this approach is that the most advanced A.I. systems are deep neural networks that "learn" how to do things on their own, in ways that aren't always interpretable by humans. We can glean some kinds of information from their internal structure, but only in limited ways, at least for the moment. This is the black box problem of A.I. So even if we had a full and exact rubric of consciousness, it would be difficult to apply it to the machines we use every day.

And the authors of the recent report are quick to note that theirs is not a definitive list of what makes one conscious. They rely on an account of "computational functionalism," according to which consciousness is reduced to pieces of information passed back and forth within a system, like in a pinball machine. In principle, according to this view, a pinball machine could be conscious, if it were made much more complex. (That might mean it's not a pinball machine anymore; let's cross that bridge if we come to it.) But others have proposed theories that take our biological or physical features, social or cultural contexts, as essential pieces of consciousness. It's hard to see how these things could be coded into a machine.

And even to researchers who are largely on board with computational functionalism, no existing theory seems sufficient for consciousness.

"For any of the conclusions of the report to be meaningful, the theories have to be correct," said Dr. Lindsay. "Which they're not." This might just be the best we can do for now, she added.

After all, does it seem like any one of these features, or all of them combined, comprise what William James described as the "warmth" of conscious experience? Or, in Thomas Nagel's words, "what it is like" to be you? There is a gap between the ways we can measure subjective experience with science and subjective experience itself. This is what David Chalmers has labeled the "hard problem" of consciousness. Even if an A.I. system has recurrent processing, a global workspace, and a sense of its physical location — what if it still lacks the thing that makes it *feel like* something?

When I brought up this emptiness to Robert Long, a philosopher at the Center for A.I. Safety who led work on the report, he said, "That feeling is kind of a thing that happens whenever you try to scientifically explain, or reduce to physical processes, some high-level concept."

The stakes are high, he added; advances in A.I. and machine learning are coming faster than our ability to explain what's going on. In 2022, Blake Lemoine, an engineer at Google, argued that the company's LaMDA chatbot was conscious (although most experts disagreed); the further integration of generative A.I. into our lives means the topic may become more contentious. Dr. Long argues that we have to start making some claims about what might be conscious and bemoans the "vague and sensationalist" way we've gone about it, often conflating subjective experience with general intelligence or rationality. "This is an issue we face right now, and over the next few years," he said.

As Megan Peters, a neuroscientist at the University of California, Irvine, and an author of the report, put it, "Whether there's somebody in there or not makes a big difference on how we treat it."

We do this kind of research already with animals, requiring careful study to make the most basic claim that other species have experiences similar to our own, or even understandable to us. This can resemble a fun house activity, like shooting empirical arrows from moving platforms toward shape-shifting targets, with bows that occasionally turn out to be spaghetti. But sometimes we get a hit. As Peter Godfrey-Smith wrote in his book "Metazoa," cephalopods probably have a robust but categorically different kind of subjective experience from humans. Octopuses have something like 40 million neurons in each arm. What's that like?

We rely on a series of observations, inferences and experiments — both organized and not — to solve this problem of other minds. We talk, touch, play, hypothesize, prod, control, X-ray and dissect, but, ultimately, we still don't know what makes us conscious. We just know that we are.

Oliver Whang is a writer based in Brooklyn. He started writing for The Times in 2020. More about Oliver Whang

A version of this article appears in print on , Section D, Page 3 of the New York edition with the headline: Is Your A.I. Conscious? Follow This Checklist.