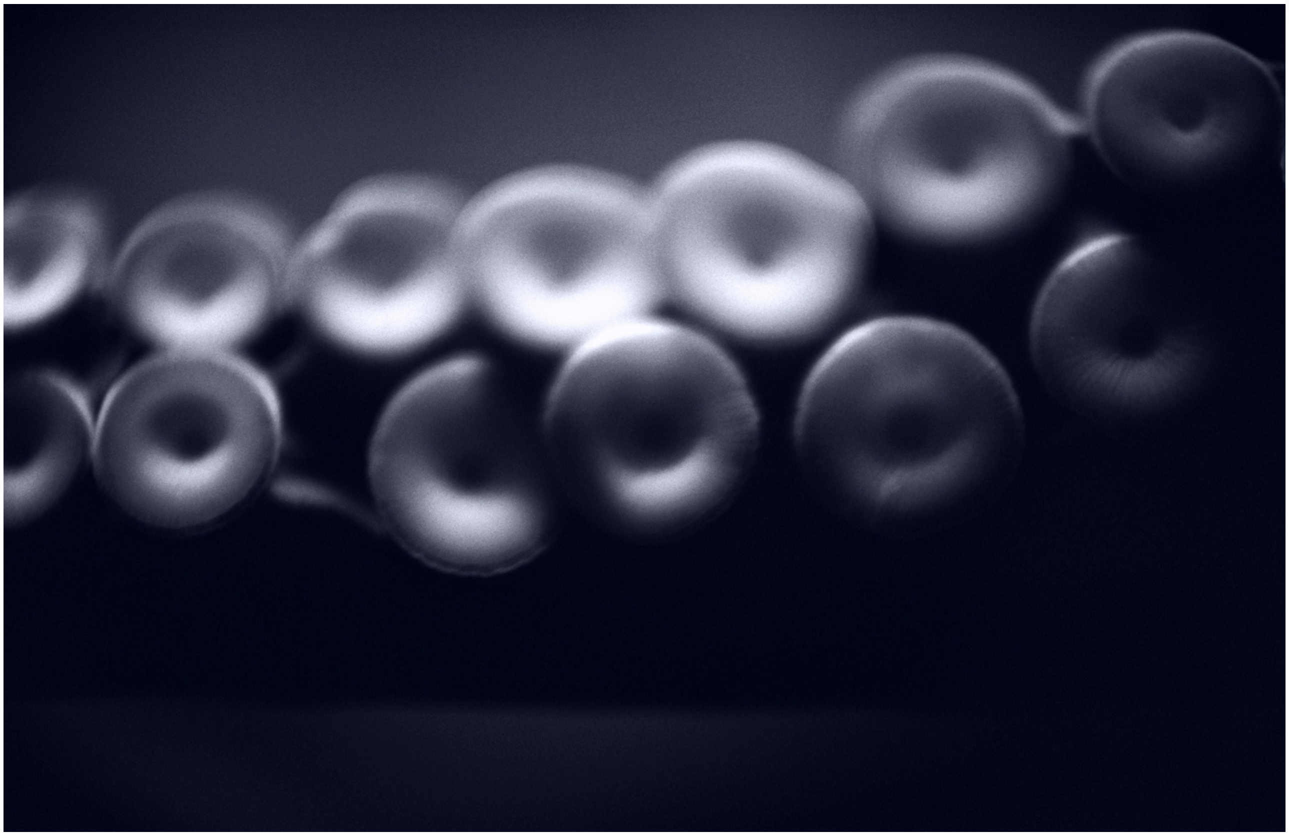


HENRY HORENSTEIN/GETTY

ARTIFICIAL INTELLIGENCE

What an octopus's mind can teach us about AI's ultimate mystery

Machine consciousness has been debated—and dismissed—since Turing. Yet it still shapes our thinking about AIs like GPT-3.

by **Will Douglas Heaven**

August 25, 2021

August 25, 2021

Machines with minds are mainstays of science fiction—the idea of a robot that somehow replicates consciousness through its hardware or software has been around so long it feels familiar.

Such machines don't exist, of course, and maybe never will. Indeed, the concept of a machine with a subjective experience of the world and a first-person view of itself goes against the grain of mainstream AI research. It collides with questions about the nature of consciousness and self—things we still don’t entirely understand. Even imagining such a machine's existence raises serious ethical questions that we may never be able to answer. What rights would such a being have, and how might we safeguard them? And yet, while conscious machines may still be mythical, their very possibility shapes how we think about the machines we are building today.

We can imagine what it would be like to observe the world through a kind of sonar. But that’s still not what it must be like for a bat, with its bat mind.

HENRY HORENSTEIN/GETTY

As Christof Koch, a neuroscientist studying consciousness, has put it: “We know of no fundamental law or principle operating in this universe that forbids the existence of subjective feelings in artifacts designed or evolved by humans.”

In my late teens I used to enjoy turning people into zombies. I’d look into the eyes of someone I was talking to and fixate on the fact that their pupils were not black dots but holes. When it came, the effect was instantly disorienting, like switching between images in an optical illusion. Eyes stopped being windows onto a soul and became hollow balls. The magic gone, I’d watch the mouth of whoever I was talking to open and close robotically, feeling a kind of mental vertigo.

The impression of a mindless automaton never lasted long. But it brought home the fact that what goes on inside other people's heads is forever out of reach. No matter how strong my conviction that other people are just like me—with conscious minds at work behind the scenes, looking out through those eyes, feeling hopeful or tired—impressions are all we have to go on. Everything else is guesswork.

Alan Turing understood this. When the mathematician and computer scientist asked the question "Can machines think?" he focused exclusively on outward signs of thinking—what we call intelligence. He proposed answering by playing a game in which a machine tries to pass as a human. Any machine that succeeded—by giving the impression of intelligence—could be said to have intelligence. For Turing, appearances were the only measure available.

But not everyone was prepared to disregard the invisible parts of thinking, the irreducible experience of the thing having the thoughts—or what we would call consciousness. In 1948, two years before Turing described his "Imitation Game," Geoffrey Jefferson, a pioneering brain surgeon, gave an influential speech to the Royal College of Surgeons of England about the Manchester Mark 1, a room-sized computer that the newspapers were hailing as an "electronic brain." Jefferson set a far higher bar than Turing: "Not until a machine can write a sonnet or compose a concerto because of thoughts and emotions felt, and not by the chance fall of symbols, could we agree that machine equals brain—that is, not only write it but know that it had written it."

Jefferson ruled out the possibility of a thinking machine because a machine lacked consciousness, in the sense of subjective experience and self-awareness ("pleasure at its successes, grief when its valves fuse"). Yet fast-forward 70 years and we live with Turing's legacy, not Jefferson's. It is routine to talk about intelligent machines, even though most would agree that those machines are mindless. As in the case of what philosophers call "zombies"—and as I used to like to pretend I observed in people—it is logically possible that a being can act intelligent when there is nothing going on "inside."

But intelligence and consciousness are different things: intelligence is about doing, while consciousness is about being. The history of AI has focused on the former and ignored the latter. If a machine ever did exist as a conscious being, how would we ever know? The answer is entangled with some of the biggest mysteries about how our brains—and minds—work.

One of the problems with testing a machine's apparent consciousness is that we don't really have a good idea of what it means for anything to be conscious. Emerging theories from neuroscience typically group things like attention, memory, and problem-solving as forms of "functional" consciousness: in other words, how our brains carry out the activities with which we fill our waking lives.

But there's another side to consciousness that remains mysterious. First-person, subjective experience—the feeling of being in the world—is known as "phenomenal" consciousness. Here we can group everything from sensations like pleasure and pain, to emotions like fear and anger and joy, to the peculiar private experiences of hearing a dog bark or tasting a salty pretzel or seeing a blue door.

For some, it's not possible to reduce these experiences to a purely scientific explanation. You could lay out everything there is to say about how the brain produces the sensation of tasting a pretzel—and it would still say nothing about what tasting that pretzel was actually like. This is what David Chalmers at New York University, one of the most influential philosophers studying the mind, calls "the hard problem."

> **Today's AI is nowhere close to being intelligent, never mind conscious. Even the most impressive deep neural networks are totally mindless.**

Philosophers like Chalmers suggest that consciousness cannot be explained by today's science. Understanding it may even require a new physics—perhaps one that includes a different type of stuff from which consciousness is made. Information is one candidate. Chalmers has pointed out that explanations of the universe have a lot to say about the external properties of objects and how they interact, but very little about the internal properties of those objects. A theory of consciousness might require cracking open a window into this hidden world.

In the other camp is Daniel Dennett, a philosopher and cognitive scientist at Tufts University, who says that phenomenal consciousness is simply an illusion, a story our brains create for ourselves as a way of making sense of things. Dennett does not so much explain consciousness as explain it away.

But whether consciousness is an illusion or not, neither Chalmers nor Dennett denies the possibility of conscious machines—one day.

Today's AI is nowhere close to being intelligent, never mind conscious. Even the most impressive deep neural networks—such as DeepMind's game-playing AlphaZero or large language models like OpenAI's GPT-3—are totally mindless.

Yet, as Turing predicted, people often refer to these AIs as intelligent machines, or talk about them as if they truly understood the world—simply because they can appear to do so.

Frustrated by this hype, Emily Bender, a linguist at the University of Washington, has developed a thought experiment she calls the octopus test.

In it, two people are shipwrecked on neighboring islands but find a way to pass messages back and forth via a rope slung between them. Unknown to them, an octopus spots the messages and starts examining them. Over a long period of time, the octopus learns to identify patterns in the squiggles it sees passing back and forth. At some point, it decides to intercept the notes and, using what it has learned of the patterns, begins to write squiggles back by guessing which squiggles should follow the ones it received.

If the humans on the islands do not notice and believe that they are still communicating with one another, can we say that the octopus understands language? (Bender's octopus is of course a stand-in for an AI like GPT-3.) Some might argue that the octopus does understand language here. But Bender goes on: imagine that one of the islanders sends a message with instructions for how to build a coconut catapult and a request for ways to improve it.

An AI acting alone might benefit from a sense of itself in relation to the world. But machines cooperating as a swarm may perform better by experiencing themselves as parts of a group rather than as individuals.

HENRY HORENSTEIN/GETTY

What does the octopus do? It has learned which squiggles follow other squiggles well enough to mimic human communication, but it has no idea what the squiggle "coconut" on this new note really means. What if one islander then asks the other to help her defend herself from an attacking bear? What would the octopus have to do to continue tricking the islander into thinking she was still talking to her neighbor?

The point of the example is to reveal how shallow today's cutting-edge AI language models really are. There is a lot of hype about natural-language processing, says Bender. But that word "processing" hides a mechanistic truth.

Humans are active listeners; we create meaning where there is none, or none intended. It is not that the octopus's utterances make sense, but rather that the islander can make sense of them, Bender says.

For all their sophistication, today's AIs are intelligent in the same way a calculator might be said to be intelligent: they are both machines designed to convert input into output in ways that humans—who have minds—choose to interpret as meaningful. While neural networks may be loosely modeled on brains, the very best of them are vastly less complex than a mouse's brain.

And yet, we know that brains can produce what we understand to be consciousness. If we can eventually figure out how brains do it, and reproduce that mechanism in an artificial device, then a conscious machine might still be possible.

When I try to imagine the world from the point of view of a machine, I can only draw on what consciousness means to me. My conception of a conscious machine is undeniably—perhaps unavoidably—human-like. It is the only form of consciousness I can imagine, as it is the only one I have experienced. But is that really what it would be like to be an AI?

It's probably hubristic to think so. The project of building intelligent machines is biased toward human intelligence. But the animal world is filled with a vast range of possible alternatives, from birds to bees to cephalopods.

A few hundred years ago the accepted view, pushed by René Descartes, was that only humans were conscious. Animals, lacking souls, were seen as mindless robots. Few think that today: if we are conscious, then there is little reason not to believe that mammals, with their similar brains, are conscious too. And why draw the line around mammals? Birds appear to reflect when they solve puzzles. Most animals, even invertebrates like shrimp and lobsters, show signs of feeling pain, which would suggest they have some degree of subjective experience.

But how can we truly picture what that must feel like? As the philosopher Thomas Nagel noted, it must "be like" something to be a bat, but what that is we cannot even imagine—because we have no idea what it would be like to observe the world through a kind of sonar. We can imagine what it might be like for a human to do this (perhaps by closing our eyes and picturing a sort of echolocation point cloud of our surroundings), but that's not what it must be like for a bat, with its bat mind.

Another way of approaching the question is by considering cephalopods, especially octopuses. These animals are known to be smart and curious—it's no coincidence Bender used them to make her point. But they have a very different kind of intelligence that evolved entirely separately from that of all other intelligent species. The last common ancestor that we share with an octopus was probably a tiny worm-like creature that lived 600 million years ago. Since then, the myriad forms of vertebrate life—fish, reptiles, birds, and mammals among them—have developed their own kinds of mind along one branch, while cephalopods developed another.

It's no surprise, then, that the octopus brain is quite different from our own. Instead of a single lump of neurons governing the animal like a central control unit, an octopus has multiple brain-like organs that seem to control each arm separately. For all practical purposes, these creatures are as close to an alien intelligence as anything we are likely to meet. And yet Peter Godfrey-Smith, a philosopher who studies the evolution of minds, says that when you come face to face with a curious cephalopod, there is no doubt there is a conscious being looking back.

A few hundred years ago the accepted view was that only humans were conscious. Animals, lacking souls, were seen as mindless robots. Few think that today.

In humans, a sense of self that persists over time forms the bedrock of our subjective experience. We are the same person we were this morning and last week and two years ago, back as far as we can remember. We recall places we visited, things we did. This kind of first-person outlook allows us to see ourselves as agents interacting with an external world that has other agents in it—we understand that we are a thing that does stuff and has stuff done to it. Whether octopuses, much less other animals, think that way isn't clear.

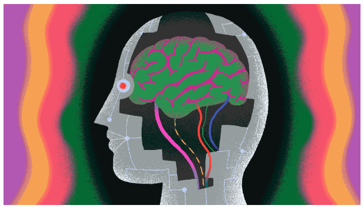
In a similar way, we cannot be sure if having a sense of self in relation to the world is a prerequisite for being a conscious machine. Machines cooperating as a swarm may perform better by experiencing themselves as parts of a group than as individuals, for example. At any rate, if a potentially conscious machine were ever to exist, we'd run into the same problem assessing whether it really was conscious that we do when trying to determine intelligence: as Turing suggested, defining intelligence requires an intelligent observer. In other words, the intelligence we see in today's machines is projected on them by us—in a very similar way that we project meaning onto messages written by Bender's octopus or GPT-3. The same will be true for consciousness: we may claim to see it, but only the machines will know for sure.

If AIs ever do gain consciousness (and we take their word for it), we will have important decisions to make. We will have to consider whether their subjective experience includes the ability to suffer pain, boredom, depression, loneliness, or any other unpleasant sensation or emotion. We might decide a degree of suffering is acceptable, depending on whether we view these AIs more like livestock or humans.

Some researchers who are concerned about the dangers of super-intelligent machines have suggested that we should confine these AIs to a virtual world, to prevent them from manipulating the real world directly. If we believed them to have human-like consciousness, would they have a right to know that we'd cordoned them off into a simulation?

Others have argued that it would be immoral to turn off a conscious machine, that this would be akin to ending a life. There are related scenarios, too. Would it be ethical to retrain a conscious machine if it meant deleting its memories? Could we copy that AI without harming its sense of self? What if consciousness turned out to be useful during training, when subjective experience helped the AI learn, but was a hindrance when running a trained model? Would it be okay to switch consciousness on and off?

Related Story



Artificial general intelligence: Are we close, and does it even make sense to try?

A machine that could think like a person has been the guiding vision of AI research since the earliest days—and remains its most divisive idea.

This only scratches the surface of the ethical problems. Many researchers, including Dennett, think that we shouldn't try to make conscious machines even if we can. The philosopher Thomas Metzinger has gone as far as calling for a moratorium on work that could lead to consciousness, even if it isn't the intended goal.

If we decided that conscious machines had rights, would they also have responsibilities? Could an AI be expected to behave ethically itself, and would we punish it if it didn't? These questions push into yet more thorny territory, raising problems about free will and the nature of choice. Animals have conscious experiences and we allow them certain rights, but they do not have responsibilities. Still, these boundaries shift over time. With conscious machines, we can expect entirely new boundaries to be drawn.

It's possible that one day there could be as many forms of consciousness as there are types of AI. But we will never know what it is like to be these machines, any more than we know what it is like to be an octopus or a bat or even another person. There may be forms of consciousness we don't recognize for what they are because they are so radically different from what we are used to.

Faced with such possibilities, we will have to choose to live with uncertainties.

And we may decide that we're happier with zombies. As Dennett has argued, we want our AIs to be tools, not colleagues. "You can turn them off, you can tear them apart, the same way you can with an automobile," he says. "And that's the way we should keep it."

by Will Douglas Heaven

Will Douglas Heaven is a senior editor for AI at MIT Technology Review.

MAGAZINE

The Mind issue

This story was part of our September/October 2021 issue.

Explore the issue →

DEEP DIVE

ARTIFICIAL INTELLIGENCE