What Would It Mean for AI to Become Conscious?

By **Vanessa Bates Ramirez** - Mar 26, 2019

As artificial intelligence systems take on more tasks and solve more problems, it's hard to say which is rising faster: our interest in them or our fear of them. Futurist Ray Kurzweil famously predicted that "By 2029, computers will have emotional intelligence and be convincing as people."

We don't know how accurate this prediction will turn out to be. Even if it takes more than 10 years, though, is it really possible for machines to become conscious? If the machines Kurzweil describes say they're conscious, does that mean they actually are?

Perhaps a more relevant question at this juncture is: what *is* consciousness, and how do we replicate it if we don't understand it?

In a panel discussion at South By Southwest titled "How AI Will Design the Human Future," experts from academia and industry discussed these questions and more.

Wait, What Is AI?

Most of AI's recent feats—diagnosing illnesses, participating in debate, writing realistic text—involve machine learning, which uses statistics to find patterns in large datasets then uses those patterns to make predictions. However, "AI" has been used to refer to everything from basic software automation and algorithms to advanced machine learning and deep learning.

"The term 'artificial intelligence' is thrown around constantly and often incorrectly," said Jennifer Strong, a reporter at the *Wall Street Journal* and host of the podcast "The Future of Everything." Indeed, one study found that 40 percent of European companies that claim to be working on or using AI don't actually use it at all.

Dr. Peter Stone, associate chair of computer science at UT Austin, was the study panel chair on the 2016 One Hundred Year Study on Artificial Intelligence (or AI100) report. Based out of Stanford University, AI100 is studying and anticipating how AI will impact our work, our cities, and our lives.

"One of the first things we had to do was define AI," Stone said. They defined it as a collection of different technologies inspired by the human brain to be able to perceive their surrounding environment and figure out what actions to take given these inputs.

Modeling on the Unknown

Here's the crazy thing about that definition (and about AI itself): we're essentially trying to re-create the abilities of the human brain without having anything close to a thorough understanding of how the human brain works.

"We're starting to pair our brains with computers, but brains don't understand computers and computers don't understand brains," Stone said. Dr. Heather Berlin, cognitive neuroscientist and professor of psychiatry at the Icahn School of Medicine at Mount Sinai, agreed. "It's still one of the greatest mysteries how this three-pound piece of matter can give us all our subjective experiences, thoughts, and emotions," she said.

This isn't to say we're not making progress; there have been significant neuroscience breakthroughs in recent years. "This has been the stuff of science

fiction for a long time, but now there's active work being done in this area," said Amir Husain, CEO and founder of Austin-based AI company Spark Cognition.

Advances in brain-machine interfaces show just how much more we understand the brain now than we did even a few years ago. Neural implants are being used to restore communication or movement capabilities in people who've been impaired by injury or illness. Scientists have been able to transfer signals from the brain to prosthetic limbs and stimulate specific circuits in the brain to treat conditions like Parkinson's, PTSD, and depression.

But much of the brain's inner workings remain a deep, dark mystery—one that will have to be further solved if we're ever to get from narrow AI, which refers to systems that can perform specific tasks and is where the technology stands today, to artificial general intelligence, or systems that possess the same intelligence level and learning capabilities as humans.

The biggest question that arises here, and one that's become a popular theme across stories and films, is if machines achieve human-level general intelligence, does that also mean they'd be conscious?

Wait, What Is Consciousness?

As valuable as the knowledge we've accumulated about the brain is, it seems like nothing more than a collection of disparate facts when we try to put it all together to understand consciousness.

"If you can replace one neuron with a silicon chip that can do the same function, then replace another neuron, and another—at what point are you still you?" Berlin asked. "These systems will be able to pass the Turing test, so we're going to need another concept of how to measure consciousness."

Is consciousness a measurable phenomenon, though? Rather than progressing by degrees or moving through some gray area, isn't it pretty black and white—a being is either conscious or it isn't?

This may be an outmoded way of thinking, according to Berlin. "It used to be that only philosophers could study consciousness, but now we can study it from a

scientific perspective," she said. "We can measure changes in neural pathways. It's subjective, but depends on reportability."

She described three levels of consciousness: pure subjective experience ("Look, the sky is blue"), awareness of one's own subjective experience ("Oh, it's me that's seeing the blue sky"), and relating one subjective experience to another ("The blue sky reminds me of a blue ocean").

"These subjective states exist all the way down the animal kingdom. As humans we have a sense of self that gives us another depth to that experience, but it's not necessary for pure sensation," Berlin said.

Husain took this definition a few steps farther. "It's this self-awareness, this idea that I exist separate from everything else and that I can model myself," he said. "Human brains have a wonderful simulator. They can propose a course of action virtually, in their minds, and see how things play out. The ability to include yourself as an actor means you're running a computation on the idea of yourself."

Most of the decisions we make involve envisioning different outcomes, thinking about how each outcome would affect us, and choosing which outcome we'd most prefer.

"Complex tasks you want to achieve in the world are tied to your ability to foresee the future, at least based on some mental model," Husain said. "With that view, I as an AI practitioner don't see a problem implementing that type of consciousness."

Moving Forward Cautiously (But Not too Cautiously)

To be clear, we're nowhere near machines achieving artificial general intelligence or consciousness, and whether a "conscious machine" is possible—not to mention necessary or desirable—is still very much up for debate.

As machine intelligence continues to advance, though, we'll need to walk the line between progress and risk management carefully.

Improving the transparency and explainability of AI systems is one crucial goal AI developers and researchers are zeroing in on. Especially in applications that could mean the difference between life and death, AI shouldn't advance without people being able to trace how it's making decisions and reaching conclusions.

Medicine is a prime example. "There are already advances that could save lives, but they're not being used because they're not trusted by doctors and nurses," said Stone. "We need to make sure there's transparency." Demanding too much transparency would also be a mistake, though, because it will hinder the development of systems that could at best save lives and at worst improve efficiency and free up doctors to have more face time with patients.

Similarly, self-driving cars have great potential to reduce deaths from traffic fatalities. But even though humans cause thousands of deadly crashes every day, we're terrified by the idea of self-driving cars that are anything less than perfect. "If we only accept autonomous cars when there's zero probability of an accident, then we will never accept them," Stone said. "Yet we give 16-year-olds the chance to take a road test with no idea what's going on in their brains."

This brings us back to the fact that, in building tech modeled after the human brain—which has evolved over millions of years—we're working towards an end whose means we don't fully comprehend, be it something as basic as choosing when to brake or accelerate or something as complex as measuring consciousness.

"We shouldn't charge ahead and do things just because we can," Stone said. "The technology can be very powerful, which is exciting, but we have to consider its implications."

Image Credit: agsandrew / Shutterstock.com

VANESSA BATES RAMIREZ

Vanessa is senior editor of Singularity Hub. She's interested in renewable energy, health and medicine, international development, and countless other topics. When she's not reading or writing you can usually find her outdoors, in water, or on a plane.