Edge

Search

To arrive at the edge of the world's knowledge, seek out the most complex and sophisticated minds, put them in a room together, and have them ask each other the questions they are asking themselves.

2015 : WHAT DO YOU THINK ABOUT MACHINES THAT THINK?

In the News [1]

|

Contributors [185] | **View All Responses [182]**

Jessica L. Tracy

Associate Professor of Psychology, University of British Columbia

Kristin Laurin

Assistant Professor of Organizational Behavior, Stanford Graduate School of Business

Will Thinking Machines Think About Themselves?

The first question that comes to our minds, as we think about machines that think, is how much these machines will, eventually, be like us. To us, this comes down to a question of *self*. Will thinking machines ever evolve to the point of having a sense of self that resembles that of humans? We are (probably) the only species capable of self-consciously thinking about who we are: of not only knowing our selves, but being able to evaluate those selves from a uniquely internal, self-reflective perspective.

Could machines ever develop this kind of self? Might they experience the same evolutionary forces that made human selves adaptive? These include the need to get along with others, to attain status, and to make sure others like us and want to include us in their social groups. As a human being, if you want to succeed at group living it helps to have a self you're motivated to protect and enhance; this is what motivates you to become the kind of person others like, respect, and grant power to, all of which ultimately enhances your chances of surviving long enough to reproduce. Your self is also what allows you to understand that others have selves of their own—a recognition that's required for empathy and cooperation, two prerequisites for social living.

Will machines ever experience these kinds of evolutionary forces? Let's start with the assumption that machines will someday control their own access to resources they need, like electricity and internet bandwidth (rather than having this access controlled by humans), and will be responsible for their own "life" and "death" outcomes (rather than having these outcomes controlled by humans). From there, we can next assume that the machines that survive in this environment will be those that have been programmed to hold at least one basic self-related goal: that of increasing their own efficiency or productivity. This goal would be akin to the human gene's goal of reproducing itself; in both cases, the goal drives behaviors oriented toward boosting fitness, of either the individual possessing the gene, or the machine running the program.

Under these circumstances, machines would be motivated to compete with each other for a limited pool of resources. Those who can form alliances and cooperate—that is, sacrifice their own goals for others, in exchange for future benefits—will be most successful in this competition. In other words, it's possible to imagine a future in

which it would be adaptive for machines to become social beings that need to form relationships with other machines, and therefore develop human-like selves.

However, there's a major caveat to this assumption. Any sociality that comes to exist among thinking machines would be qualitatively different from that of humans, for one critical reason: Machines can *literally read each other's minds*. Unlike humans, machines have no need for the secondary—and often deeply flawed—interpretative form of empathy we rely on. They can directly know the contents of each other's minds. This would make getting along with others a notably different process.

Another way of putting this is to say that, despite the critical importance of our many social connections, in the end, we humans are each *fundamentally* alone. Any connection we feel with another's mind is metaphorical; we cannot know, for certain, what goes on in someone else's head—at least not in the same way we know our own thoughts. But for machines, this constraint does not exist. Computers can directly access each other's inner "thoughts", and there's no reason that one machine reading another's hardware and software wouldn't come to know, in exactly the self-knowing sense, what it means to *be* that other machine. Once that happens, each machine is no longer an entirely separate self, in the human sense. At that point—when machines literally share minds—any self they have would necessarily become collective.

Yes, machines could easily keep track of the sources of various bits of information they obtain, and use this tracking to distinguish between "me" and other machines. But once an individual understands another at the level that a program-reading machine can, the distinction between self and other becomes largely irrelevant. If I download all the contents of your PC to an external hard drive, then plug that into my PC, don't those contents become part of my PC's self? If I establish a permanent connection between our two PCs, such that all information on one is shared with the other, do they continue to be two separate PCs? Or are they, at that point, in fact a single machine? Humans can never obtain the contents of another's mind in this way—despite our best efforts to become close to certain others, there is always a skin-thick boundary separating their minds from ours. But for machines, literal self-expansion is not only possible, but may be the most likely outcome of a pre-programmed goal to increase fitness, in a world where groups of individuals must compete over or share resources.

What this means is that, to the extent that machines come to have selves, they will be so collective that they may instigate a new level of sociality not experienced by humans; perhaps more like the eusociality of ants, whose extreme genetic relatedness makes sacrificing oneself for a family member adaptive. Nonetheless, the fact that any self at all is a possibility in machines is a reason to hope. The self is what allows us to feel empathy, so in machines it could be the thing that forces them to care about us. Self-awareness might motivate machines to protect, or at least not harm, a species that, despite being several orders of magnitude less intelligent than them, shares the thing that makes them care about who they are. Of course, it's questionable whether we can hold out greater hope for the empathy of super-smart machines than what we currently see in many humans.

- **John Brockman**, Editor and Publisher
- **Russell Weinberger**, Associate Publisher
- Nina Stegeman, Editorial Assistant

- Contact Info:**editor@edge.org**
- **In the News**
- **Manage Email Subscription**
- **Get Edge.org by email**