# Artificial intelligence has learned to probe the minds of other computers

August 06, 2018

Anyone who's had a frustrating interaction with Siri or Alexa knows that digital assistants just don't get humans. What they need is what psychologists call theory of mind, an awareness of others' beliefs and desires. Now, computer scientists have created an artificial intelligence (AI) that can probe the "minds" of other computers and predict their actions, the first step to fluid collaboration among machines—and between machines and people.

"Theory of mind is clearly a crucial ability," for navigating a world full of other minds says Alison Gopnik, a developmental psychologist at the University of California, Berkeley, who was not involved in the work. By about the age of 4, human children understand that the beliefs of another person may diverge from reality, and that those beliefs can be used to predict the person's future behavior. Some of today's computers can label facial expressions such as "happy" or "angry"—a skill associated with theory of mind—but they have little understanding of human emotions or what motivates us.

The new project began as an attempt to get humans to understand computers. Many algorithms used by AI aren't fully written by programmers, but instead rely on the machine "learning" as it sequentially tackles problems. The resulting computer-generated solutions are often black boxes, with algorithms too complex for human insight to penetrate. So Neil Rabinowitz, a research scientist at DeepMind in London, and colleagues created a theory of mind AI called "ToMnet" and had it observe other AIs to see what it could learn about how they work.