THIS WEEK 28 February 2018

DeepMind AI is learning to understand the ‘thoughts’ of others

The firm’s new artificial intelligence has developed a theory of mind, passing an important psychological assessment that most children only develop around age 4

AIs are beginning to comprehend that others see things differently
Curtis Johnson/Aurora/getty

By **Timothy Revell**

MACHINES are getting to know each other better. An artificial intelligence, developed by Google-owned research firm DeepMind, can now pass an important psychological assessment that most children only develop the skills to pass at around age 4. Its aptitude in this key theory of mind test may lead to AIs that are more human-like.

Most humans regularly think about other people’s desires, beliefs or intentions. For a long time, this was thought to be uniquely human, but an increasing body of somewhat controversial evidence suggests that some other animals, such as chimps, bonobos, orangutans and ravens may have theory of mind (see “What do you think?“). However, the idea that machines could share these abilities is normally reserved for sci-fi.

DeepMind thinks otherwise. The firm created its latest AI with the intention of it developing a basic theory of mind. The AI is called Theory of Mind-net, or ToM-net for short. In a virtual world, ToM-net is able to not just predict what other AI agents will do, but also understand that they may hold false beliefs about the world.

For humans, the idea that others can hold false beliefs seems very natural, especially if you follow politics closely, or read the comment section on news websites. However, humans don’t actually understand that other people can hold false beliefs until around age 4. “It’s a classic developmental stage for young children,” says Peter Stone at the University of Texas at Austin.

One of the main reasons we know about this is a psychology experiment called the Sally-Anne test. In the test, Anne watches Sally leave an object somewhere, only for it to be moved without Sally seeing. Anne, who has seen everything, is then asked where Sally will first look for the object. To pass the test, Anne needs to be able to distinguish between where the object actually is and where Sally thinks it is. In other words, Anne needs to understand that Sally may hold a false belief about the object’s location.

Guess what I’m doing

To mimic this set-up for AIs, ToM-net plays the role of Anne in a virtual world consisting of an 11-by-11 grid, some internal walls and four objects. A different AI agent also inhabits the gridworld and is set a task, unknown to ToM-net, about walking to one of the four objects. The agent is rewarded depending on how optimal its path is. ToM-net has to predict what is going to happen.

In a similar way to the Sally-Anne test, the DeepMind team switches some objects during the experiment. For example, the agent might see its preferred blue object in one location, but would be told to walk to another object first. While concentrating on this sub-goal, the preferred object would be moved and the agent may or may not see this happen, depending on its position.

“The AI can predict others’ behaviour, and figure out when they have false beliefs about the world”

Surprisingly, ToM-net is able to accurately predict and understand what this agent and others also used are trying to do, essentially passing this form of the Sally-Anne test and exhibiting some basic theory of mind. “It can learn the differences between agents, predict how they might behave differently, and figure out when agents will have false beliefs about the world,” says Neil Rabinowitz at DeepMind.

This is a big step. Making computer programs that mimic behaviours like theory of mind could improve our understanding of people and other animals, says Christopher Lucas at the University of Edinburgh, UK.

But Alan Wagner at Georgia Tech Research Institute says the 11-by-11 grid set-up is too simplistic for the researchers to claim they have captured the idea of theory of mind.

Outside of the debate as to whether ToM-net truly exhibits theory of mind, there is a possibility it might help make more human-like AIs. “I wouldn’t be surprised if this can make things like chatbots seem a lot like humans,” says Joanna Bryson at the University of Bath in the UK.

In turn, this may make interactions smoother. “The more our machines can learn to understand others, the better they can interpret requests, help find information, explain what they’re doing, teach us new things and tailor their responses to individuals,” says Rabinowitz.

What do you think?

It is difficult to know if animals have theory of mind because we can’t ask them what they are thinking. Instead, we use simple tests.

Mirror Test

A dot is placed on an animal’s face. If it recognises itself in a mirror, it will try to wipe the mark off, demonstrating basic self-recognition – a part of theory of mind. Children do this from about 15 months. Great apes, dolphins, killer whales and Eurasian magpies have passed this test as well.

Joint Attention

The ability to both guide and follow someone else’s gaze is not unique to humans, but it is part of theory of mind. Children learn to do it in their first year or so. Many other primates, as well as dogs and horses, have the ability too.

False belief

Understanding that others can hold false beliefs is perhaps the trickiest test to pass. Children develop this skill around age 4. Some apes and birds may have it too, as well as a new artificial intelligence agent.

This article appeared in print under the headline “Theory of a machine’s mind”

Magazine issue 3167, due to be published 3 March 2018