Google DeepMind wants to define what counts as artificial general intelligence

technologyreview.com/2023/11/16/1083498/google-deepmind-what-is-artificial-general-intelligence-agi

Will Douglas Heaven

AGI, or artificial general intelligence, is one of the hottest topics in tech today. It's also one of the most controversial. A big part of the problem is that few people agree on what the term even means. Now a team of Google DeepMind researchers has put out a paper that cuts through the cross talk with not just one new definition for AGI but a whole taxonomy of them.

To come up with the new definition, the Google DeepMind team started with prominent existing definitions of AGI and drew out what they believe to be their essential common features.

The team also outlines five ascending levels of AGI: emerging (which in their view includes cutting-edge chatbots like ChatGPT and Bard), competent, expert, virtuoso, and superhuman (performing a wide range of tasks better than all humans, including tasks humans cannot do at all, such as decoding other people's thoughts, predicting future events, and talking to animals). They note that no level beyond emerging AGI has been achieved.

"This provides some much-needed clarity on the topic," says Julian Togelius, an AI researcher at New York University, who was not involved in the work. "Too many people sling around the term AGI without having thought much about what they mean."

The researchers posted their paper online last week with zero fanfare. In an exclusive conversation with two team members—Shane Legg, one of DeepMind's co-founders, now billed as the company's chief AGI scientist, and Meredith Ringel Morris, Google DeepMind's principal scientist for human and AI interaction—I got the lowdown on why they came up with these definitions and what they wanted to achieve.

A sharper definition

"I see so many discussions where people seem to be using the term to mean different things, and that leads to all sorts of confusion," says Legg, who came up with the term in the first place around 20 years ago. "Now that AGI is becoming such an important topic—you know, even the UK prime minister is talking about it—we need to sharpen up what we mean."

It wasn't always this way. Talk of AGI was once derided in serious conversation as vague at best and magical thinking at worst. But buoyed by the hype around generative models, buzz about AGI is now everywhere.

When Legg suggested the term to his former colleague and fellow researcher Ben Goertzel for the title of Goertzel's 2007book about future developments in AI, the hand-waviness was kind of the point. "I didn't have an especially clear definition. I didn't really feel it was necessary," says Legg. "I was actually thinking of it more as a field of study, rather than an artifact."

His aim at the time was to distinguish existing AI that could do one task very well, like IBM's chess-playing program Deep Blue, from hypothetical AI that he and many others imagined would one day do many tasks very well. Human intelligence is not like Deep Blue, says Legg: "It is a very broad thing."

But over the years, people started to think of AGI as a potential property that actual computer programs might have. Today it's normal for top AI companies like Google DeepMind and OpenAI to make bold public statements about their mission to build such programs.

"If you start having those conversations, you need to be a lot more specific about what you mean," says Legg.

For example, the DeepMind researchers state that an AGI must be both general-purpose and high-achieving, not just one or the other. "Separating breadth and depth in this way is very useful," says Togelius. "It shows why the very accomplished AI systems we've seen in the past don't qualify as AGI."

They also state that an AGI must not only be able to do a range of tasks, it must also be able to learn how to do those tasks, assess its performance, and ask for assistance when needed. And they state that what an AGI can do matters more than how it does it.

It's not that the way an AGI works doesn't matter, says Morris. The problem is that we don't know enough yet about the way cutting-edge models, such as large language models, work under the hood to make this a focus of the definition.

"As we gain more insights into these underlying processes, it may be important to revisit our definition of AGI," says Morris. "We need to focus on what we can measure today in a scientifically agreed-upon way."

Measuring up

Measuring the performance of today's models is already controversial, with researchers debating what it really means for a large language model to pass dozens of high school tests and more. Is it a sign of intelligence? Or a kind of rote learning?

Assessing the performance of future models that are even more capable will be more difficult still. The researchers suggest that if AGI is ever developed, its capabilities should be evaluated on an ongoing basis, rather than through a handful of one-off tests.

The team also points out that AGI does not imply autonomy. “There’s often an implicit assumption that people would want a system to operate completely autonomously,” says Morris. But that’s not always the case. In theory, it’s possible to build super-smart machines that are fully controlled by humans.

One question the researchers don’t address in their discussion of *what* AGI is, is *why* we should build it. Some computer scientists, such as Timnit Gebru, founder of the Distributed AI Research Institute, have argued that the whole endeavor is weird. In a talk in April on what she sees as the false (even dangerous) promise of utopia through AGI, Gebru noted that the hypothetical technology “sounds like an unscoped system with the apparent goal of trying to do everything for everyone under any environment.”

Most engineering projects have well-scoped goals. The mission to build AGI does not. Even Google DeepMind’s definitions allow for AGI that is indefinitely broad and indefinitely smart. “Don’t attempt to build a god,” Gebru said.

In the race to build bigger and better systems, few will heed such advice. Either way, some clarity around a long-confused concept is welcome. “Just having silly conversations is kind of uninteresting,” says Legg. “There’s plenty of good stuff to dig into if we can get past these definition issues.”