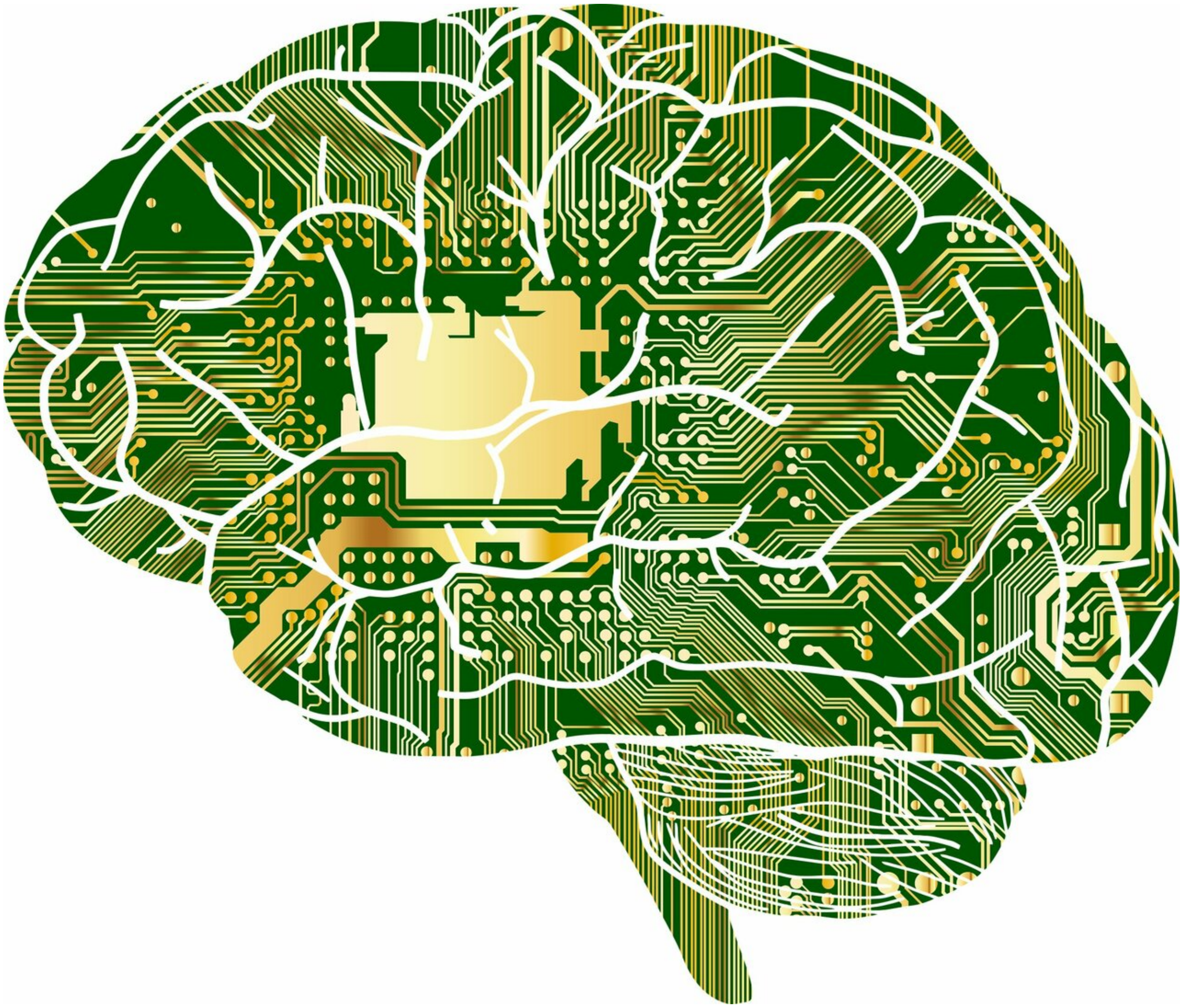
Mathematical paradoxes demonstrate the limits of AI

March 17 2022



Credit: CC0 Public Domain

Humans are usually pretty good at recognizing when they get things wrong, but artificial intelligence systems are not. According to a new study, AI generally suffers from inherent limitations due to a century-old mathematical paradox.

Like some people, AI systems often have a degree of confidence that far exceeds their actual abilities. And like an overconfident person, many AI systems don't know when they're making mistakes. Sometimes it's even more difficult for an AI system to realize when it's making a mistake than to produce a correct result.

Researchers from the University of Cambridge and the University of Oslo say that instability is the Achilles' heel of modern AI and that a mathematical paradox shows AI's limitations. Neural networks, the state of the art tool in AI, roughly mimic the links between neurons in the brain. The researchers show that there are problems where stable and accurate neural networks exist, yet no algorithm can produce such a network. Only in specific cases can algorithms compute stable and accurate neural networks.

The researchers propose a classification theory describing when neural networks can be trained to provide a trustworthy AI system under certain specific conditions. Their results are reported in the *Proceedings of the National Academy of Sciences.*

Deep learning, the leading AI technology for pattern recognition, has been the subject of numerous breathless headlines. Examples include diagnosing disease more accurately than physicians or preventing road accidents through autonomous driving. However, many deep learning systems are untrustworthy and easy to fool.

"Many AI systems are unstable, and it's becoming a major liability, especially as they are increasingly used in high-risk areas such as disease diagnosis or autonomous vehicles," said co-author Professor Anders Hansen from Cambridge's Department of Applied Mathematics and Theoretical Physics. "If AI systems are used in areas where they can do real harm if they go wrong, trust in those systems has got to be the top priority."

The paradox identified by the researchers traces back to two 20th century mathematical giants: Alan Turing and Kurt Gödel. At the beginning of the 20th century, mathematicians attempted to justify mathematics as the ultimate consistent language of science. However, Turing and Gödel showed a paradox at the heart of mathematics: it is impossible to prove whether certain mathematical statements are true or false, and some computational problems cannot be tackled with algorithms. And, whenever a mathematical system is rich enough to describe the arithmetic we learn at school, it cannot prove its own consistency.

Decades later, the mathematician Steve Smale proposed a list of 18 unsolved mathematical problems for the 21st century. The 18th problem concerned the limits of intelligence for both humans and machines.

"The paradox first identified by Turing and Gödel has now been brought forward into the world of AI by Smale and others," said co-author Dr. Matthew Colbrook from the Department of Applied Mathematics and Theoretical Physics. "There are fundamental limits inherent in mathematics and, similarly, AI algorithms can't exist for certain problems."

The researchers say that, because of this paradox, there are cases where good neural networks can exist, yet an inherently trustworthy one cannot be built. "No matter how accurate your data is, you can never get the perfect information to build the required neural network," said co-author Dr. Vegard Antun from the University of Oslo.

The impossibility of computing the good existing neural network is also true regardless of the amount of training data. No matter how much data an algorithm can access, it will not produce the desired network. "This is similar to Turing's argument: there are computational problems that cannot be solved regardless of computing power and runtime," said Hansen.

The researchers say that not all AI is inherently flawed, but it's only reliable in specific areas, using specific methods. "The issue is with areas where you need a guarantee, because many AI systems are a black box," said Colbrook. "It's completely fine in some situations for an AI to make mistakes, but it needs to be honest about it. And that's not what we're seeing for many systems—there's no way of knowing when they're more confident or less confident about a decision."

"Currently, AI systems can sometimes have a touch of guesswork to them," said Hansen. "You try something, and if it doesn't work, you add more stuff, hoping it works. At some point, you'll get tired of not getting what you want, and you'll try a different method. It's important to understand the limitations of different approaches. We are at the stage where the practical successes of AI are far ahead of theory and understanding. A program on understanding the foundations of AI computing is needed to bridge this gap."

"When 20th-century mathematicians identified different paradoxes, they didn't stop studying mathematics. They just had to find new paths, because they understood the limitations," said Colbrook. "For AI, it may be a case of changing paths or developing new ones to build systems that can solve problems in a trustworthy and transparent way, while understanding their limitations."

The next stage for the researchers is to combine approximation theory, numerical analysis and foundations of computations to determine which neural networks can be computed by algorithms, and which can be made stable and trustworthy. Just as the paradoxes on the limitations of mathematics and computers identified by Gödel and Turing led to rich foundation theories—describing both the limitations and the possibilities of mathematics and computations—perhaps a similar foundations theory may blossom in AI.

More information: Matthew J. Colbrook et al, The difficulty of computing stable and accurate neural networks: On the barriers of deep learning and Smale's 18th problem, *Proceedings of the National Academy of Sciences* (2022). DOI: 10.1073/pnas.2107151119

Douglas Heaven, Why deep-learning AIs are so easy to fool, *Nature* (2019). DOI: 10.1038/d41586-019-03013-5

Provided by University of Cambridge

Citation: Mathematical paradoxes demonstrate the limits of AI (2022, March 17) retrieved 23 April 2024 from https://techxplore.com/news/2022-03-mathematical-paradoxes-limits-ai.html