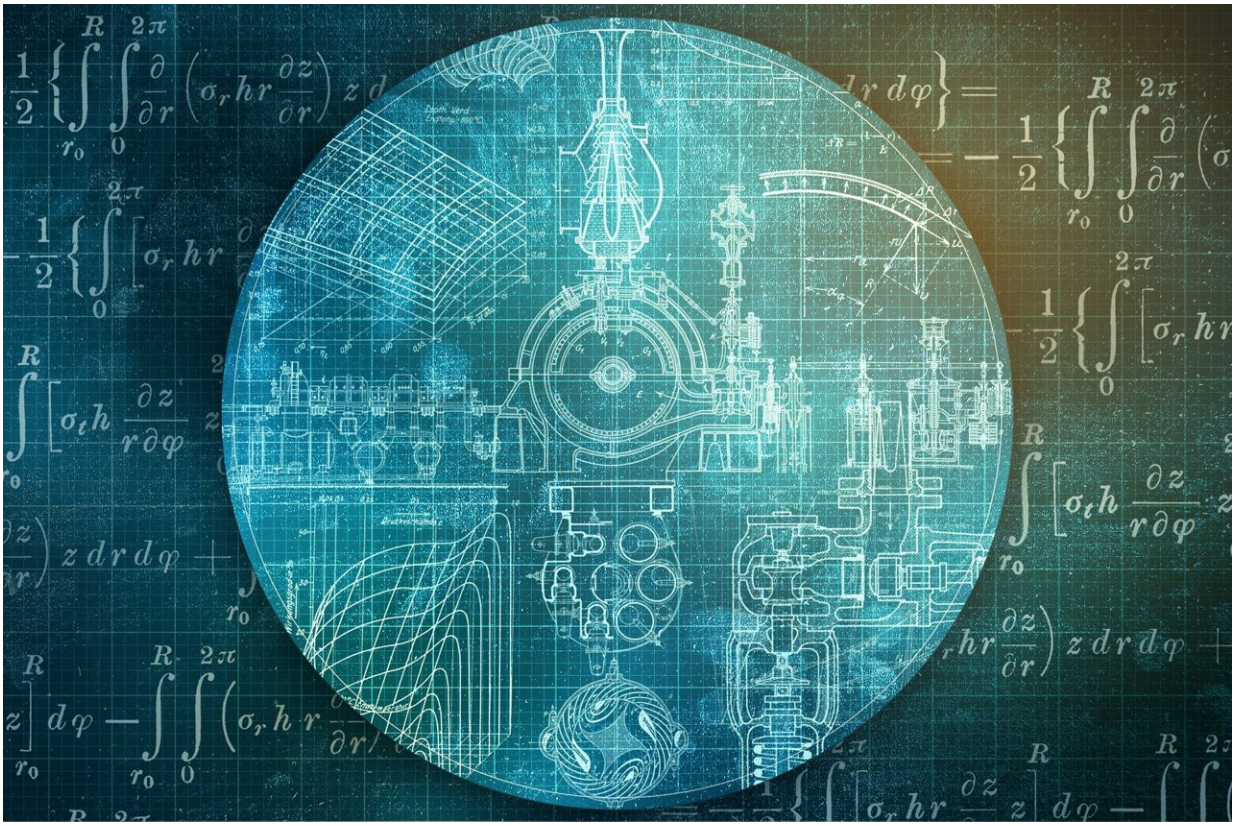
Study pinpoints the weaknesses in AI

January 11 2024



Credit: Pixabay/CC0 Public Domain

ChatGPT and other solutions built on machine learning are surging. But even the most successful algorithms have limitations. Researchers from University of Copenhagen have proven mathematically that apart from simple problems it is not possible to create algorithms for AI that will always be stable. The study, posted to the *arXiv* preprint server, may lead

to guidelines on how to better test algorithms and reminds us that machines do not have human intelligence after all.

Machines interpret medical scanning images more accurately than doctors, they translate foreign languages, and may soon be able to drive cars more safely than humans. However, even the best algorithms do have weaknesses. A research team at Department of Computer Science, University of Copenhagen, tries to reveal them.

Take an automated vehicle reading a road sign as an example. If someone has placed a sticker on the sign, this will not distract a human driver. But a machine may easily be put off because the sign is now different from the ones it was trained on.

"We would like algorithms to be stable in the sense, that if the input is changed slightly the output will remain almost the same. Real life involves all kinds of noise which humans are used to ignore, while machines can get confused," says Professor Amir Yehudayoff, heading the group.

A language for discussing weaknesses

"I would like to note that we have not worked directly on automated car applications. Still, this seems like a problem too complex for algorithms to always be stable," says Yehudayoff, adding that this does not necessarily imply major consequences in relation to development of automated cars. "If the algorithm only errs under a few very rare circumstances this may well be acceptable. But if it does so under a large collection of circumstances, it is bad news."

The scientific article cannot be applied by industry for identifying bugs in its algorithms. This wasn't the intention, the professor explains. "We are developing a language for discussing the weaknesses in machine

learning algorithms. This may lead to development of guidelines that describe how algorithms should be tested. And in the long run this may again lead to development of better and more stable algorithms."

From intuition to mathematics

A possible application could be for testing algorithms for protection of digital privacy.

"Some company might claim to have developed an absolutely secure solution for privacy protection. Firstly, our methodology might help to establish that the solution cannot be absolutely secure. Secondly, it will be able to pinpoint points of weakness," says Yehudayoff.

First and foremost, though, the scientific article contributes to theory. Especially the mathematical content is groundbreaking, he adds,

"We understand intuitively, that a stable algorithm should work almost as well as before when exposed to a small amount of input noise. Just like the road sign with a sticker on it. But as theoretical computer scientists we need a firm definition. We must be able to describe the problem in the language of mathematics. Exactly how much noise must the algorithm be able to withstand, and how close to the original output should the output be if we are to accept the algorithm to be stable? This is what we have suggested an answer to."

Important to keep limitations in mind

The scientific article has received large interest from colleagues in the theoretical computer science world, but not from the tech industry. Not yet at least.

"You should always expect some delay between a new theoretical development and interest from people working in applications," says Yehudayoff. "And some theoretical developments will remain unnoticed forever."

However, he does not see that happening in this case:

"Machine learning continues to progress rapidly, and it is important to remember that even solutions which are very successful in the real world still do have limitations. The machines may sometimes seem to be able to think but after all they do not possess human intelligence. This is important to keep in mind."

More information: Zachary Chase et al, Replicability and stability in learning, *arXiv* (2023). DOI: 10.48550/arxiv.2304.03757

Provided by University of Copenhagen