August 29, 2022

Maths, Physics & Chem

Mathematical paradoxes unearth the boundaries of AI

by Matthew J. Colbrook[1,2,3] | Postdoctoral research fellow, Vegard Antun[1,2,3] | Postdoctoral research fellow, Anders C. Hansen[1,2,3] | Professor

[1]: Centre Sciences des Données, École Normale Supérieure and Department of Applied Mathematics and Theoretical Physics, University of Cambridge
[2]: Department of Mathematics, University of Oslo
[3]: Department of Applied Mathematics and Theoretical Physics, University of Cambridge, and Department of Mathematics, University of Oslo

This Break was edited by Sofia Spataro, *Senior Scientific Editor* - TheScienceBreaker

ABSTRACT

Instability is AI's Achilles' heel. We show the following paradox: there are cases where stable and accurate AI exists, but it can never be trained by any algorithm. We initiate a foundations theory for when AI can be trained - such a programme will shape political and legal decision-making in the coming decades, and have a significant impact on markets for AI technologies.

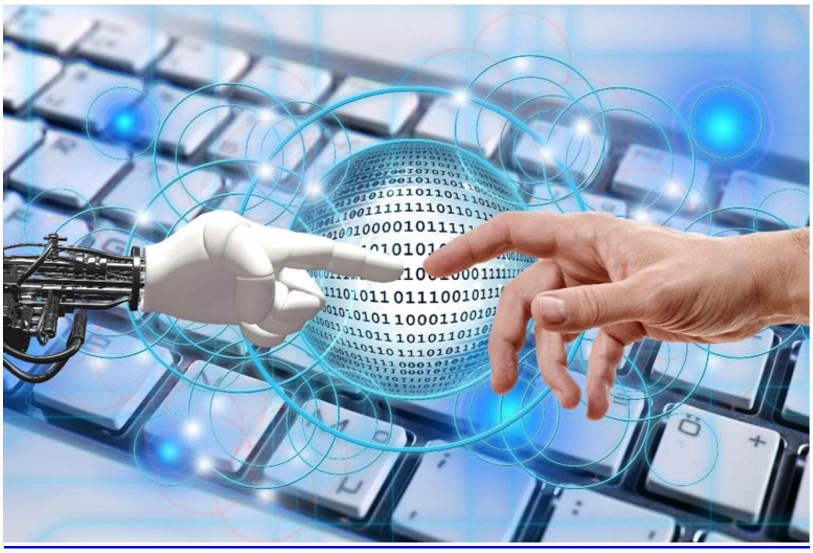

Image credits: Image by Gerd Altmann from Pixabay

Artificial intelligence (AI) has made monumental strides in numerous areas over the last decade. One only has to look at the news to see the latest breakthrough, whether it's an AI beating a world-champion player at some game, achieving human-level object recognition, or diagnosing cancer from medical scans. However, there is another side to this story. It is also becoming increasingly apparent that many AI systems are non-robust and unstable to tiny changes in the input data. The AI may even hallucinate and produce nonsensical output with high prediction confidence. Good examples of hallucinations can be seen in the results of Facebook and NYU's fastMRI challenge (this is a competition for AI magnetic resonance imaging, MRI, reconstruction). These issues are a severe concern in safety-critical applications, such as medical diagnosis and self-driving cars, and a serious concern within legal frameworks for the use of AI. Alarmingly, these complications also seem to occur even for problems where we know that classical methods produce stable and thus safe solutions. Instability appears to be the Achilles' heel of modern AI.

Neural networks (NNs) are the current state-of-the-art tool in AI and are motivated by the links between neurons in the brain. The "universal approximation theorem" says that stable problems can be solved stably with a NN. Therefore, we are led to the following puzzling question: why does AI lead to unstable methods and AI-generated hallucinations, even in scenarios where one can prove that stable and accurate NNs exist? In our work, we seek to answer this question. We show that there are problems where stable and accurate NNs exist, yet no algorithm can produce such a network. Regardless of how many computational resources or data one throws at the problem, this impossibility result holds. Moreover, whether it is possible also depends on the accuracy one wants. Only in specific cases do training algorithms exist for stable and accurate NNs. We also propose a mathematical classification theory describing when NNs can be trained to provide a trustworthy AI system. For example, under suitable assumptions and using a new optimisation technique, we explicitly construct NNs that are provably stable (and robust against so-called "adversarial attacks") and accurate for MRI imaging. To prove our results, we use a framework called the Solvability Complexity Index, which allows us to classify the difficulty of mathematical computational problems and prove that algorithms are optimal.

The above results show an essential difference between abstract existence and trainability. Mathematically proving the existence of a good NN is not enough - one must also show that it can be obtained in practice. This paradox is very much related to the work of Alan Turing and Kurt Gödel. About 100 years ago, mathematicians set out to show that mathematics was the ultimate consistent language of the universe. There was a tremendous amount of optimism, similar to the optimism we see in AI today. However, Turing and Gödel turned this optimism on its head: it is impossible to prove whether certain mathematical statements are true or false, and some problems cannot be tackled with algorithms. Much later, the mathematician Steve Smale proposed a list of 18 unsolved mathematical problems for the 21st century. His 18th problem, featured in the title of our paper, concerned the limits of intelligence for both humans and machines. The mathematics of foundations, i.e., figuring out what is and is not possible, is now entering the world of AI.

The above paradox may seem gloomy, but it is important to stress that not all AI is inherently flawed. The above results show that AI is only reliable in specific areas, using specific methods. The problem now becomes figuring out these cases. 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. Currently, the practical successes of AI are far ahead of our understanding of these systems. A programme on the foundations of AI is needed to bridge this gap. Figuring out what can and cannot be done will be healthy for AI in the long run. The paradoxes on the limitations of mathematics and computers identified by Gödel and Turing led to rich foundation theories, new techniques, and methodology. Perhaps a similar foundations theory may blossom in AI.