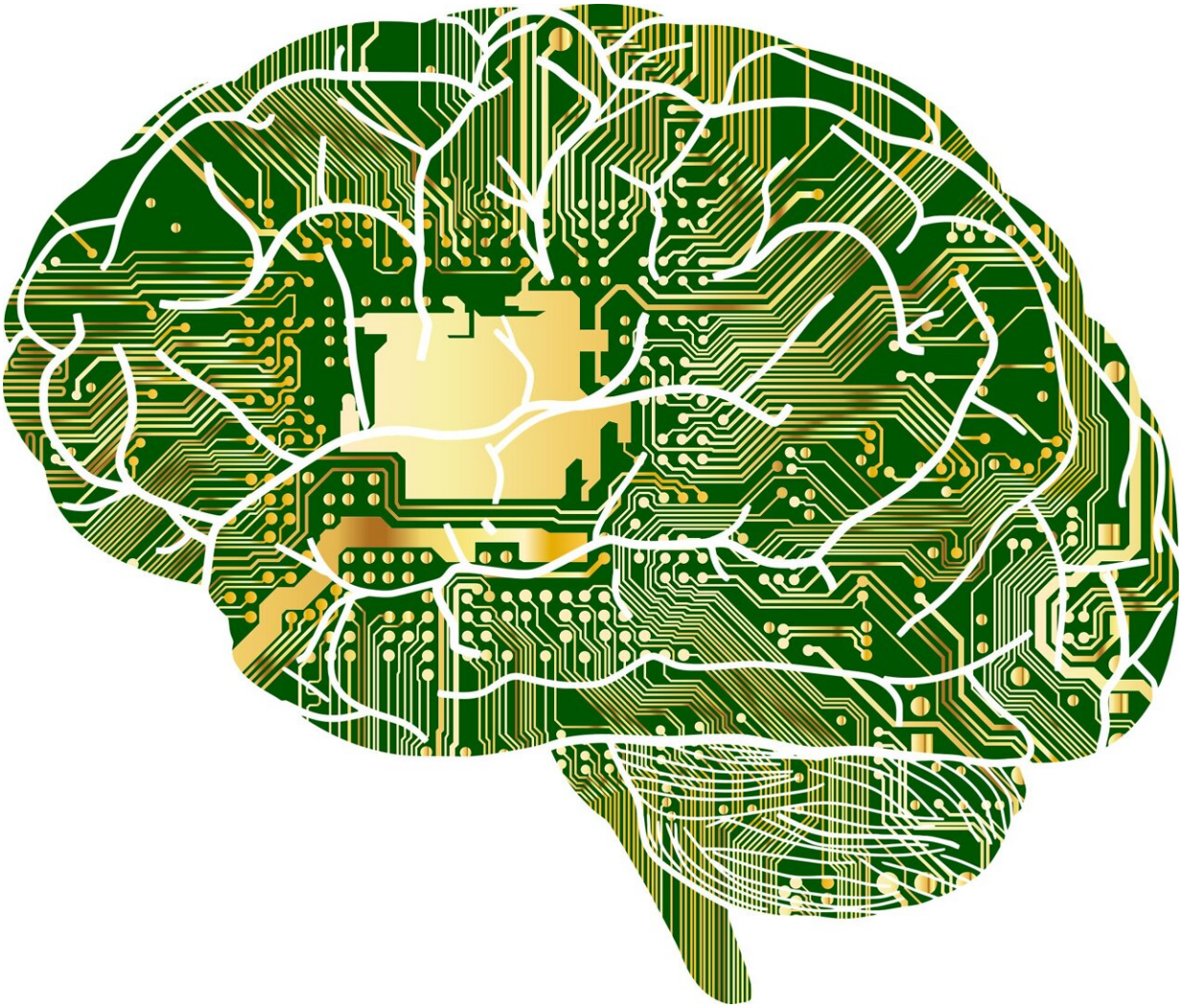
Researchers psychoanalyse AI

September 21 2021, by Jesper Spangsmark Nielsen, Hanne Kokkegård



Credit: Pixabay/CC0 Public Domain

We do not know exactly what is going on inside the "brain" of artificial

intelligence (AI), and therefore we are not able to accurately predict its actions. We can run tests and experiments, but we cannot always predict and understand why AI does what it does.

Just like humans the development of artificial intelligence is based on experiences (in the form of data when it comes to AI). That is why the way artificial intelligence acts sometimes catch us by surprise, and there are countless examples of artificial intelligence behaving sexist, racist, or just inappropriate.

"Just because we can develop an algorithm that lets artificial intelligence find patterns in data to best solve a task, it does not mean that we understand what patterns it finds. So even though we have created it, it does not mean that we know it, "says Professor Søren Hauberg, DTU Compute.

A paradox called the black box problem. Which on the one hand is rooted in the self-learning nature of artificial intelligence and on the other hand, in that the fact that so far it has not been possible to look into the "brain" of AI and see what it does with the data to form the basis of its learning.

If we could find out what data AI works with and how, it would correspond to something in between exams and psychoanalysis—in other words, a systematic way to get to know artificial intelligence much better. So far it has just not been possible, but now Søren Hauberg and his colleagues have developed a method based on classical geometry, which makes it possible to see how an artificial intelligence has formed it's "personality."

Messy brain

It requires very large data sets to teach robots to grab, throw, push, pull,

walk, jump, open doors and etc., and artificial intelligence only uses the data that enables it to solve a specific task. The way artificial intelligence sorts out useful from useless data, and ultimately sees the patterns on which it subsequently bases its actions, is by compressing its data into neural networks.

However, just like when we humans pack things together, it can easily look messy to others, and it can be hard to figure out which system we have used.

For example, if we pack our home together with the purpose that it should be as compact as possible, then a pillow easily ends up in the soup pot to save space. There is nothing wrong with that, but outsiders could easily draw the wrong conclusion; that pillows and soup pots were something we had intended to use together. And that has been the case so far when we humans tried to understand what systematics artificial intelligence works by. According to Søren Hauberg, however, it is now a thing of the past:

"In our basic research, we have found a systematic solution to theoretically go backwards, so that we can keep track of which patterns are rooted in reality and which have been invented by compression. When we can separate the two, we as humans can gain a better understanding of how artificial intelligence works, but also make sure that the AI does not listen to false patterns. "

Søren and his DTU colleagues have drawn on mathematics developed in the 18th century for used to draw maps. These classic geometric models have found new applications in machine learning, where they can be used to make a map of how compression has moved data around and thus go backwards through the AI's neural network and understand the learning process.

Gives back control

In many cases, the industry refrains from using artificial intelligence, specifically in those parts of production where safety is a crucial parameter. Fear losing control of the system, so that accidents or errors occur if the algorithm encounters situations that it does not recognize and has to take action itself.

The new research gives back some of the lost control and understanding. Making it more likely that we will apply AI and machine learning to areas that we do not do today.

"Admittedly, there is still some of the unexplained part left, because part of the system has arisen from the model itself finding a pattern in data. We can not verify that the patterns are the best, but we can see if they are sensible. That is a huge step toward more confidence in the AI, "says Søren Hauberg.

The mathematical method was developed together with the Karlsruhe Institute of Technology and the industrial group Bosch Center for Artificial Intelligence in Germany. The latter has implemented software from DTU in its robot algorithms. The results have just been published in an article presented at the Robotics: Science and Systems conference.

More information: Hadi Beik-Mohammadi et al, Learning Riemannian Manifolds for Geodesic Motion Skills (2021), arXiv:2106.04315v2 [cs.RO], arxiv.org/abs/2106.04315

Provided by Technical University of Denmark