The interconnected neurons of the brain serve as a model for artificial neural networks, which a team at the Max Planck Institute for Mathematics in the Sciences is researching.

PHOTO: SVEN DÖRING FOR MPG

INTELLIGENCE WITH A PLAN

TEXT: THOMAS BRANDSTETTER

Self-learning algorithms are turning our society upside down. But all too frequently, even their developers do not fully understand how they work. Researchers at the Max Planck Institute for Mathematics in the Sciences want to remedy the situation with fundamental theories of machine learning.

PHOTO: REED HUTCHINSON / UCLA

Guido Montúfar conducts research at the Max Planck Institute for Mathematics in the Sciences and the University of California in Los Angeles.

Learning algorithms are taking over more and more tasks that, until recently, only humans seemed capable of. No online service can do without them, even if they are simply used to place advertisements. Artificial intelligence, or AI for short, can translate texts or even write them itself. One such example is the chatbot ChatGPT, which assembles astonishingly meaningful answers to almost any question from countless texts available on the Internet. Artificial intelligence is also trying its hand at art and, when installed in driver assistance systems, becomes a more careful road user than the person behind the wheel. The current boom in AI is primarily explained by the continuing advancement of artificial neural networks, which are loosely modeled on the human brain. The functioning of the brain is simulated on computers as deeply branched networks of artificial neurons, whose connections dynamically adapt to new experiences in order to recognize patterns in data or learn new behaviors. And just as with the brain, it is often difficult to understand exactly what is happening in its electronic equivalents. While it is true that their success serves as proof of their validity and, for many purposes, it may be sufficient to merely view them as a useful black box, experts believe it is necessary to provide AIs with a sound theoretical foundation in order to further exploit the technology's enormous potential. This would allow them to get a detailed understanding of how the algorithms learn.

Guido Montúfar and his team at the Max Planck Institute for Mathematics in the Sciences and the University of California, Los Angeles, are also working toward this goal.

"We're investigating the mathematical side of artificial neural networks," the researcher explains. According to Montúfar, much of their development has taken place on a practical level, while theories have often been lacking. "People have tried a lot of things," he says. "Sometimes their intuitions have been spot-on and sometimes less so." Together with his colleagues, the mathematician is now working on a theory of neural networks.

Large networks even for small amounts of data

In essence, the current triumph of artificial intelligence can be attributed to three factors. The first factor is that developers have access to a constantly increasing amount of computing power in the form of improved hardware located in ever larger data cen-

→

ters. This allows ever larger neural networks to be created, in which countless artificial neurons are connected today. Just as the number and strength of the synapses – the connections between the nerve cells – are important in the brain, the connections in an artificial neural network are also crucial. For example, GPT-3, a precursor of ChatGPT, is said to have 175 billion connections – the human brain has 100 trillion. The second factor is the increased availability of data for training artificial neural networks thanks to advancing digitalization, which has given AI a significant boost. And third, international cooperation in the development of new algorithms is making software ever more complex as a result. "One surprise in AI development, for example, was that large neural networks can produce meaningful results even with relatively little training data," Montúfar explains. In the past, he continues, it was believed that the high flexibility of large networks was only beneficial when using correspondingly large data sets. This is due to the fact that huge neural networks are by definition exceedingly complex. And accordingly, he says, this could easily lead to poor decisions when training is implemented with only a small amount of data. "This is an expectation which is actually mathematically sound," explains Guido Montúfar. "But the theory on which this expectation was based was incomplete." Even without theory, he says, it was soon observed in practice that the larger neural networks are, the better they function.

SUMMARY

In many cases, we simply have no idea how artificial neural networks learn to solve a task and what criteria they use to do so. This can be problematic when artificial intelligence is applied in medicine or image recognition for autonomous driving, for example.

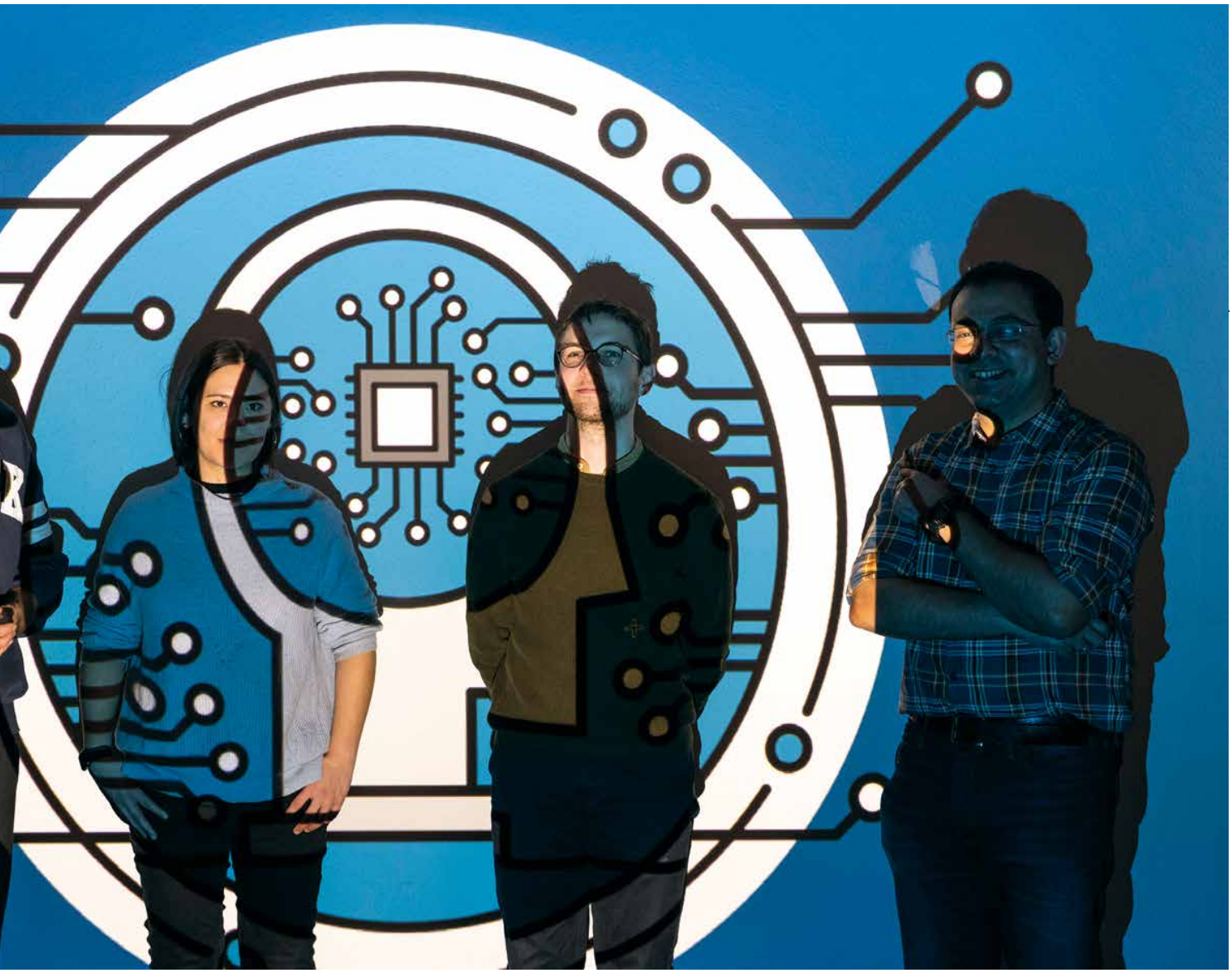
A theory of artificial neural networks would help to make their decision-making comprehensible. It could also speed up the search for suitable algorithms.

Neural network training identifies mathematical functions that solve a problem. Normally, many functions are suitable for this, but not all are equally good. To find the optimal function, a theory should also consider other properties of the functions beyond their ability to solve the training problem.

So you could very well argue that the reasons why a network works are not that important – the main thing is that it works. "Ten years ago, quite a lot of people were convinced that you didn't need much theory," Montúfar says. He thinks this may have been because a theoretical understanding truly isn't needed in some cases. Developing a mathematical theory isn't for everyone, even for people who successfully program AI algorithms. But if you have no theory to build on, you simply have to try out numerous different ideas. This is a laborious process that consumes time and resources. Montúfar gives the example of a chemist who has no idea about elements and molecules and mixes different substances together at random just to see what happens. On top of that, AI could increasingly be used in sensitive areas, such as to support medical diagnoses or autonomously drive vehicles. In these cases, it would be useful to know exactly how it reaches decisions.

There is also the fact that, despite all the successes of AI, there is certainly still room for improvement. For example, the dangers posed by malicious attacks, such as in the area of autonomous driving, are particularly worrying. These attacks could trick the AI system that recognizes and interprets traffic signs, and cause accidents as a result. Even a minimal but targeted change to a road sign can, in some circumstances, cause the system to believe that the speed limit is 120 miles-per-hour instead of 50. To the human eye, by contrast, the change to the road sign would be barely perceptible. But why is it possible to put an artificial neural network on such a fatally wrong track? "To answer this question and ultimately prevent such

PHOTO: SVEN DÖRING FOR MPG



Illuminating artificial intelligence: Johannes Müller, Marie-Charlotte Brandenburg, Pierre Bréchet, and Pradeep Kumar Banerjee (from left) are working on a theory of artificial neural networks to help better understand how AI works.

attacks, we need a deeper understanding of how networks function," Montúfar says.

Understanding how to improve data protection

In addition, there is still much to be desired when it comes to protecting privacy. For example, machines designed to predict whether and how quickly a patient will recover in a hospital are pre-trained using data from real patients. "As a result, this data is in the system in one form or another," says Montúfar. "And a customer who buys this system could try to open it up and read this private data." To guarantee that something like this is not possible, he says, you would have to know very precisely what goes on inside such a system. Currently, the Mathematical Machine Learning group is looking at the phenomenon in which neural networks prefer to find solutions with certain mathematical properties. "Even if there are several options, ultimately the network will usually choose solutions with certain characteristics," Montúfar says. "We're trying to characterize exactly what those preferences are." The researcher illustrates the interrelationship by comparing the search for a suitable solution to a hike through a landscape. The network, or rather the algorithm

→

GRAPHIC: GUIDO MONTÚFAR / MPI FOR MATHEMATICS IN THE SCIENCES

Sample solutions: the individual charts correspond to different configurations of an example neural network that is supposed to identify tables in pictures. The greater the number of colored areas in a chart, the more complex the associated function the network uses to solve the task. The Leipzig-based mathematicians are investigating the conditions on which the complexity of the determined function depends.

with which it works, begins its search at a starting point chosen more or less arbitrarily by the developer. Because of this, it tends to seek out solutions that are close to this point. In a sense, the neural network is trapped in the immediate vicinity of the nearby solution. "We now want to understand what the neighborhood around this solution looks like and why the neural network can't get out of it," Montúfar explains.

Artificial intelligence could be tasked, for example, with establishing a connection between data on living space and rental prices and finding a suitable mathematical function for this purpose. A reasonable solution would not simply consist of a line or, more precisely, a mathematical function that provides the best possible connection between the individual data points. In order to show a plausible dependence of continuously increasing rent with increasing apartment size, this line should also be as smooth as possible, i.e., it should not have any kinks. "Hence, a suitable area to start searching the landscape of solutions would be around functions that are very smooth," Montúfar explains. "Conversely, it would be absurd to look for lines that are totally jagged." In order to stop choosing the starting point for the search completely arbitrarily and instead systematically select suitable areas for it, the researchers want to know what the landscape of solutions looks like. "We have already made a lot of progress in this area," Montúfar says. "We are now working on a precise mathematical theory that allows concrete predictions about the behavior of a neural network under different conditions."

A theory for faster training

Another factor is that theories of complex phenomena are sometimes oversimplified, and important aspects of the actual facts are lost. For example, mathematical theories are often less complicated if the artificial neural networks they describe are considered infinite for the sake of simplicity. But even though there is indeed a trend toward ever larger networks, in reality these are of course still finite. "Such simplifications can cause big problems," Montúfar says. The reason for this, he explains, is that unforeseen effects then occur in the networks, which are huge but still finite, and these effects can be very significant. "If we could understand and explain these effects, we could train these AI systems much faster in practice," Montúfar estimates. And this would not only save a lot of electricity

and money, but also time. After all, the supercomputers that run these algorithms often operate for weeks or even months at a time.

Another important focus for Montúfar's team is the data itself, which is used to train the machines. In a similar manner to living brains, its quality has significant influence on the development of an artificial neural network. One example of this is provided by research conducted in the 1960s. It studied animals with early vision impairments and discovered that the impairment was irreversible. Even after the eye problem was corrected, the animals were unable to see sharply. "If you've learned to see with blurry images, your brain can't do much with sharp images later on," Montúfar summarizes. And in computer science, the way training data is designed also affects the learning process. "We want to know what those effects are," says the researcher, "that is, what influence the selection of training data has on the development and then later the performance of a network." If artificial intelligence development continues on its current trajectory, then the issue of artificial consciousness may arise at some point. "If we assume that consciousness exists, I see no reason why it shouldn't also exist in artificial neural networks," Montúfar says pragmatically. Many researchers are already dreaming of developing artificial intelligence to the point where it is no longer limited to doing single, specialized tasks but instead gains a thorough understanding of the world. An artificial general intelligence such as this would then be able to perform a multitude of different intellectual tasks and would probably outperform humans in each of them sooner or later.

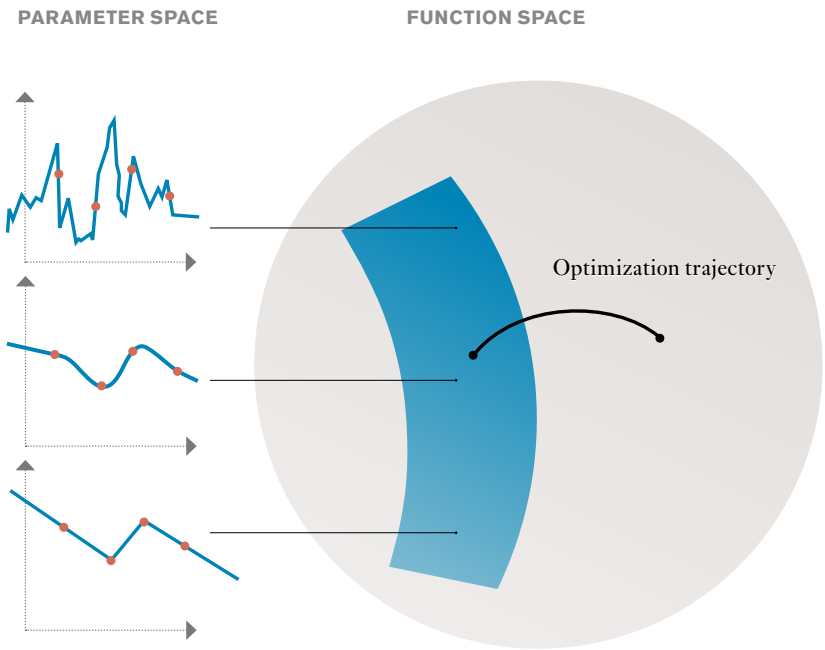
GLOSSARY

ARTIFICIAL NEURAL NETWORK is the name given to a computer-simulated system in which numerous artificial neurons are networked in a similar way to the nerve cells of the brain. Artificial neurons generate an output value from an input value and pass this on to another neuron. In neural network training, the connections between neurons are made and strengthened to optimally solve a task. These connections correspond to the criteria specified in the search for solutions to a problem and are weighted according to their relevance.

However, the question remains as to when and how we will decide whether AGI has been achieved. "In the past, you might have called a spam filter intelligent, but no one is impressed by this anymore," Montúfar says. So the boundary is constantly shifting, and AGI is in fact hard to define. How many tasks should this system be able to solve in order to be called "artificial general intelligence?" "Some people have already formulated tests and defined benchmarks on how to decide this," Montúfar says. "But, I think it's ultimately going to come down to practice whether people perceive a system as general intelligence or not."

"It's possible, however, that a computer intelligence will be created in a similar manner to birthing a child," Montúfar says. "Children are also intelligent beings and we don't actually have to understand anything to create them." So we might eventually create AGI without knowing exactly what we're doing. "The question is what kind of being this will be," Montúfar muses. "And if it is somehow supposed to be compatible with our society and our concepts of civilization, it would be very useful if we could understand it." But to be able to do this, he says, we must first become acquainted with the theoretical foundations of AI.

←



During learning, an artificial neural network adjusts its parameters, effectively its synaptic connections, along the optimization trajectory and seeks the blue-colored region of functions that fit all data points (red). The task is considered to have been solved very well by the function in the mid panel with as few maxima and minima as possible and without any kinks.

GRAPHIC: GCO BASED ON A TEMPLATE BY GUIDO MONTÚFAR / MPI FOR MATHEMATICS IN THE SCIENCES