3 Physics

Learning from Networks

Protein science is a field in which both practical and intellectual goals hold substantial sway. Most of the protein researchers I spoke with care deeply about both: They want to reduce the amount of human suffering caused by protein-related diseases, and they also want to satisfy their curiosity about how proteins work. At certain times, or for certain individuals, one goal will take precedence over the other, but their general coexistence means that researchers might face difficult decisions about what to prioritize, if prediction and understanding become further disentangled.

That same precarious balance between the practical and the intellectual is present across the sciences. Physics, however, is a notable exception. Though physics has been put to countless practical purposes – modern technology wouldn't exist without it – those applications tend not to play as prominent a role in the field's discourse as they might in other disciplines.

"What is the goal of physics? I think the goal of physics is to understand the universe," says Mario Krenn, a physicist and machine learning researcher at the University of Tübingen in Germany. In fields such as biology and chemistry, he says, practical goals might sometimes trump intellectual ones, but not in physics. "If you can come up with

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 12 Nov 2026 at 09:24:50, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://www.cambridge.org/core/product/0258EEDF4C0E24D8F5A87BD4DB30E244

cheaper, more robust, more efficient photovoltaic cells, who cares whether you understand it? If biochemists develop some bacteria that eat plastic and produce food, or something, no one cares [how they work], because that is amazing for humanity. In physics, it's different."

The particularly high value that physicists place on understanding, even when it has no clear link to practical goals, is nowhere clearer than in the century-long debate about what quantum mechanics tells us about the nature of reality. The mathematical basis of quantum mechanisms is incontestable: No scientist would ever dispute that the mathematical formalization of Heisenberg's uncertainty principle or the Schrödinger wave equation effectively explains a wide range of empirical observations. But much as Newton's curiosity about the reasons for the planets' motion was not entirely satisfied by Kepler's laws, the physicists of the twentieth and twenty-first centuries have agonized and argued about what quantum equations tell us about the nature of reality. How can something be both a particle and a wave, and why does the state of a particle seem to change when it is observed? These are extraordinarily difficult questions that fall far outside the boundaries of human intuition, and yet physicists continue to discuss them.

Based on that history, one might think that physicists would foreswear deep learning – what use could they have for a technology that offers prediction without understanding? But many in the field have taken the opposite approach. They are enthusiastically embracing deep learning and other AI tools will the goal of gleaning understanding from them in any way possible.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 12 Nov 2026 at 09:24:50, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://www.cambridge.org/core/product/0258EEDF4C0E24D8F5A87BD4DB30E244

Krenn is one of the central figures in this movement, though he certainly didn't start his career expecting to be. He completed his PhD at the University of Vienna, where he worked under Nobel-winning physicist Anton Zeilinger and studied a phenomenon known as quantum entanglement. Entanglement is one of those bizarre features of quantum mechanics that simply don't make sense to anyone who has lived their life and built their intuitions in a macroscopic world. When two particles get entangled, their states become linked, and that link persists no matter how far they roam from each other. They're like a pair of coins that will always both come up heads or both come up tails when they are flipped at the same time, even if they're located on opposite sides of the planet. This mystical-seeming connection is so baffling that Einstein famously referred to it as "spooky action at a distance."

In the lab, Krenn and his colleagues fired lasers through crystals to create entangled pairs of photons and then manipulated and measured those photons using mirrors, beam splitters, and other devices. Some of these experiments involved entangling more than two particles, and figuring out how to generate those complex entanglements could be a tricky task. One particular entangled state flummoxed the group for months: Try as they might, they couldn't come up with a setup of crystals, mirrors, and splitters that would create it. So Krenn wrote up a computer program that would iteratively test out different combinations of the devices they had available in the lab and evaluate whether that combination could produce the desired entanglement state. In a matter of hours, the program found a solution.[1]

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 12 Nov 2026 at 09:24:50, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://www.cambridge.org/core/product/0258EEDF4C0E24D8F5A87BD4DB30E244

Other scientists might have happily accepted that solution and moved on to conducting their experiments. But Krenn was intrigued by the program's success and curious about what he could learn from it. "That completely changed the way my career went," he says. "I never went back to the lab. I just was extremely fascinated by the possibility that the machine could teach us new stuff."

Learning from machines isn't always easy: Unlike human teachers, computer programs typically can't explain their solutions to us. So it was up to Krenn and his colleagues to study the program's solution on their own, until they could understand why it worked. The breakthrough came when Krenn uncovered an unexpected connection between the setups the program had designed and a field of mathematics called graph theory, which studies networks of nodes connected to one another by edges. He realized that all of the experimental setups he was working with could be represented as graphs, and by analyzing those graphs, he began to understand the general principles behind how and why certain setups worked.[2] This connection between quantum entanglement experiments and graph theory is still a central part of Krenn's research.

The program that taught Krenn about that connection could be considered an AI tool, though it wasn't a deep neural network. But the experience of learning so much from a machine-generated solution convinced him that AI systems, including deep neural networks, can play important roles in the quest for scientific understanding. Not long after that experience, he published a paper titled "On Scientific Understanding with Artificial Intelligence," in which he

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 12 Nov 2026 at 09:24:50, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://www.cambridge.org/core/product/0258EEDF4C0E24D8F5A87BD4DB30E244

draws on his own research, accounts that he and his coauthors gathered from other researchers, and Henk de Regt's philosophical work to argue that scientists can glean understanding from AI tools in a variety of ways.[3] Some of those approaches have already proven successful, such as investigating the predictions and solutions devised by AI, as Krenn himself did. Others are more incipient, including opening the black box of deep neural networks to sort through their inner workings and figure out how they make their predictions. Still others, such as learning completely novel scientific ideas from AI "teachers," are as yet entirely speculative.

In the paper, and in much of his work, Krenn is motivated by the idea that prediction is not enough. "An oracle that correctly predicts the outcome of every particle physics experiment, the products of every possible chemical reaction or the function of every protein would revolutionize science and technology," he writes. "However, scientists would not be entirely satisfied because they would want to comprehend how the oracle made these predictions."

Physicists like Krenn aren't content with deep neural networks that spit out predictions but do nothing to further their understanding of the universe. And as they work to squeeze scientific understanding from a recalcitrant source, their efforts could have implications that reach far beyond their own field. If physicists or other researchers eventually achieve a reliable means of explaining how a deep neural network makes its predictions, then deep neural networks could very well prove a net positive for scientific understanding. Extracting insights from large-scale datasets is difficult; training deep neural networks on those datasets and

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 12 Nov 2026 at 09:24:50, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://www.cambridge.org/core/product/0258EEDF4C0E24D8F5A87BD4DB30E244

then determining what the networks have learned could potentially be much easier.

None of this changes the fact that deep neural networks frequently dissociate prediction from understanding – and, in doing so, reduce the practical motivation to achieve understanding. A purely practical science has no time for understanding for its own sake, no matter how easy it is to achieve. But for those who adhere to the intellectual view of science, research aimed at gaining understanding from deep neural networks, if successful, could provide an invaluable boost in an increasingly prediction-focused world.

Looking for Laws

If you want to extract understanding from a deep neural network, you need to know what understanding looks like. That depends heavily on what scientific field you happen to be in. For instance, in molecular biology, scientists use a combination of mathematical formulae (such as the equation that describes an amino acid chain's free energy), models (the 3-D brass structure of lysozyme), and mechanisms (the physical process through which ATP synthase stores energy) to explain how proteins fold and function.

But physics is where math reigns supreme. The most universal and powerful mathematical formulae in physics are known as laws, and these and other equations form the bulk of a physicist's explanatory toolbox. To explain the trajectory of billiard balls after they collide, you need the laws of conservation of momentum and conservation of energy; an explanation of how changing the voltage in an electrical circuit

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 12 Nov 2026 at 09:24:50, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://www.cambridge.org/core/product/0258EEDF4C0E24D8F5A87BD4DB30E244

affects its current will invoke Ohm's law; and explaining the paths that the planets trace across the night sky requires Kepler's laws, or Newton's law of universal gravitation for added accuracy.

So physicists who want to gain understanding from deep neural networks might start by trying to extract usable equations from them. That, however, is much easier said than done. Strictly speaking, deep neural networks are just math: Information flows from neuron to neuron according to specific mathematical rules, and it's theoretically possible to write down an equation that captures everything a deep neural network does. For a sophisticated deep neural network such as AlphaFold, however, you'd need thousands of notebooks to write up such an equation in its entirety. No human could possibly gain any scientific insight from it, at least not without computer help.

The question, then, is whether it might be possible to turn all that gnarly math into something simpler. Fortunately, that's a problem very much within the remit of physics, a field that aims to capture complex physical phenomena in clean and comprehensible formulas – much as Kepler did when he developed his laws using Tycho's data. And the tools for doing so are growing more powerful.

Over the past few decades, scientists have worked to automate the search for simple mathematical laws in big, messy datasets using a machine learning technique known as symbolic regression. To apply symbolic regression to a dataset, you first need to identify the variables in that dataset that you care about, and the symbolic regression algorithm will then work to combine those variables using various

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 12 Nov 2026 at 09:24:50, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://www.cambridge.org/core/product/0258EEDF4C0E24D8F5A87BD4DB30E244

mathematical tools – addition, multiplication, exponents, trigonometric functions – until it finds an expression that effectively captures the patterns in your data.

In some cases, symbolic regression obviates the need for neural networks entirely. If there's a simple mathematical equation that describes the relationship between your input data and your output data then symbolic regression should be able to find it, and you can then use that equation to make predictions – no need to train a neural network at all. In other cases, however, neural networks can form an essential bridge between data and equation.

In a 2020 study, MIT physicists Silviu-Marian Udrescu and Max Tegmark showed that neural networks could help bring to light equations that had proven resistant to earlier symbolic regression techniques.[4] By first training a neural network on the dataset from which they were trying to extract an equation and then using that network to generate additional, simulated data, they managed to uncover new patterns that radically simplified the task of symbolic regression.

Udrescu and Tegmark didn't present their method as a strategy for understanding neural networks, but their success nevertheless presents a possible path forward for physicists who want to figure out what a trained deep neural network has learned. Their pipeline involves first getting a neural network to learn the relationships among variables in a dataset, then using that network to discover a simple equation that captures those same relationships and renders the network's predictive capacities superfluous. So if a physicist were to successfully apply that same approach to a network

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 12 Nov 2026 at 09:24:50, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://www.cambridge.org/core/product/0258EEDF4C0E24D8F5A87BD4DB30E244

that had already been trained for practical purposes, they would obtain an equation that they could use to make exactly the same predictions as the network. Unlike the network, the equation would be comprehensible.

But this recipe for obtaining concise equations from ungainly, uninterpretable neural networks is far from universal: It only works if the task on which the network was trained can be solved using a relatively simple formula. That's a huge limitation. The problems for which deep neural networks are best suited, such as protein folding, are far too complicated to be described in a single equation.

That's not to say that symbolic regression is automatically useless in such situations, however. Instead of focusing on the network's inputs and outputs, one could perform symbolic regression on particular neurons within the network to see if their activity patterns can be captured with concise formulas. In other words, you could dig into the network and get your hands dirty. That was the approach taken by New York University physicist Shirley Ho when, together with her student Miles Cranmer, she devised a new method for mixing together deep neural networks and symbolic regression.[5]

The basic idea of Ho and Cranmer's method is to use a neural network whose structure is so specialized, and so well suited to the problem at hand, that it's possible to find the specific neurons that represent a quantity you are interested in. By performing symbolic regression not on the whole network's predictions but rather on the activity patterns of those neurons, you can potentially find a formula for that quantity.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 12 Nov 2026 at 09:24:50, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://www.cambridge.org/core/product/0258EEDF4C0E24D8F5A87BD4DB30E244

Specifically, Ho, Cranmer, and their team trained networks to predict the motion of simulated objects that are attracted to each other by various forces, such as gravity. For that task, they chose a network whose structure captured these key features of the scenario's physics. One layer of the network contains specific neurons for every pair of objects in the simulation, which reflects the fact that gravitational forces exist between pairs of objects. These neurons learn to calculate those forces. The next layer contains neurons that represent individual objects, and each neuron in that layer sums the inputs from neurons in the force layer – because the total force on an object is the sum of all the individual forces it experiences.

All that structure made applying symbolic regression straightforward. Ho and Cranmer knew which neurons should learn to calculate the gravitational force between objects, so all they had to do was apply symbolic regression to those neurons' activity patterns to uncover the mathematical formula for that force. When they did so, they got back a correct mathematical equation for Newton's law of universal gravitation.

Machine learning experts call systems like this network interpretable, because it's possible to interpret, or understand, how they make their predictions. Most deep neural networks are not interpretable, so, in order to ensure that theirs would be, Ho and Cranmer had to use their knowledge of gravity to guide particular layers in the network toward taking on specific roles in the overall prediction.

Ho, Cranmer, and their collaborators built on this approach in a later study, in which they trained a network

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 12 Nov 2026 at 09:24:50, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://www.cambridge.org/core/product/0258EEDF4C0E24D8F5A87BD4DB30E244

to predict the gravitational motion not of abstract objects but of the planets and moons in our solar system.[6] To keep the network interpretable in this more realistic scenario, they modified its structure so that it would be forced to obey Newton's third law – which stipulates, for instance, that the force Mercury exerts on Venus is exactly as strong as the force Venus exerts on Mercury – and Newton's second law – which states that an object's acceleration is proportional to the total force it experiences. Here, too, they were able to interpret the network and extract Newton's law of universal gravitation.

The overall message of these studies is that forcing deep neural networks to adhere to various physical principles and laws helps ensure that they will be interpretable once trained, so that any new physics knowledge they acquire can be extracted. But imposing such constraints isn't always feasible. For some problems, especially the most challenging and interesting ones, physicists might not yet know the important background laws and principles. And it's in precisely those situations where deep neural networks, and their ability to learn to make accurate predictions with minimal guidance, are most useful.

Crystallizing Concepts

When it's successful, symbolic regression is a powerful tool. It can transform impenetrable, incomprehensible wads of data into equations, the backbone of physical understanding. But it's not a panacea. In order to apply symbolic regression to a problem, you need to know which variables matter ahead of time, since those variables will be the inputs that

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 12 Nov 2026 at 09:24:50, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://www.cambridge.org/core/product/0258EEDF4C0E24D8F5A87BD4DB30E244

you feed into the symbolic regression system. And though the important variables in physics – mass, energy, charge – might seem self-evident, they are anything but.

I started thinking about this problem seriously after interviewing Kyle Cranmer (no relation to Miles), a particle physicist at the University of Wisconsin–Madison who specializes in deep neural networks and other machine learning tools. We spoke on the afternoon of the Friday before his winter break, and we spent most of our conversation discussing the various ways that deep neural networks had already become embedded in particle physics research. As we were finishing up, however, Cranmer tacked toward a completely different subject. What had really been preoccupying him recently, he said, was the question of whether or not an AI system could learn new concepts on its own.

Cranmer had worked with Ho on some of her symbolic regression work, and so he knew that deep neural networks could learn laws. But he was acutely aware of just how much assistance the networks had needed in order to infer those laws. Through either their input and output data or their architectures, the networks had been given the concepts of position, mass, acceleration, and force. "At the beginning, you gave it this big hint: Those are the things that you should pay attention to," he says. "It's easy to take that step for granted."

Scientific concepts don't spring fully formed from the ether: They need to be articulated and refined by human scientists. Take energy, a concept that crops up everywhere in science, from Einstein's most famous equation to theories of protein folding. The law of conservation of energy, which

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 12 Nov 2026 at 09:24:50, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://www.cambridge.org/core/product/0258EEDF4C0E24D8F5A87BD4DB30E244

states that the total amount of energy in any isolated system is constant, is one of the most powerful and versatile principles in physics. But physicists such as Galileo and Newton couldn't take advantage of this law when they were devising their field-defining ideas about motion and gravity, because the modern concept of energy hadn't yet been invented.

Advancing physics requires more than just uncovering new laws. Laws are useful, but a great concept will appear in numerous different physical laws – and a really great one might even clarify the nature of reality. Several concepts proposed during the twentieth century, including quarks and dark energy, have transformed not only the academic but also the public understanding of the nature of the universe. If deep neural networks do manage to uncover a novel physical concept that stands up to scientific scrutiny, they will have made a major contribution to scientific understanding.

When we spoke, Cranmer said that he was interested in the possibility that AI systems could identify useful concepts, but he also described some challenges in the way of that vision. To make his point, he used the example of El Niño, a climate pattern characterized by unusually warm temperatures in the tropical areas of the Pacific Ocean that occurs every two to seven years. Once scientists identify that we are in an El Niño year, they know to expect droughts in some parts of the world, excessive rain in others, and more hurricanes in the Pacific than in the Atlantic. El Niño, in other words, is a useful concept. It's not a necessary one – it's possible to characterize the global climate in perfect detail without ever mentioning it – but it makes weather patterns easier for us humans to predict and understand.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 12 Nov 2026 at 09:24:50, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://www.cambridge.org/core/product/0258EEDF4C0E24D8F5A87BD4DB30E244

The question on Cranmer's mind was whether we could expect deep neural networks and other AI systems to identify useful concepts like El Niño on their own. There's good reason to think they might not. A powerful deep neural network doesn't need higher-level concepts like El Niño to make accurate predictions: It can work directly with all the innumerable details of shifting temperatures and air pressures and wind speeds. All of that information would overwhelm a human mind, which is why we devise concepts such as El Niño to help us get a handle on complex systems. Deep neural networks, however, don't have this limitation.

That observation bodes ill for our prospects of finding new scientific concepts in behemoth networks such as AlphaFold. But it also suggests a strategy for teaching deep neural networks concepts: Limit their power. If one of the layers of a network is constrained to have very few neurons – perhaps only two or three – then all of the information that the network processes will have to fit through that choke point. That means it will need to figure out some way to predict outputs on the basis of only a few variables, each of which will be learned by one of the neurons. You can't pack all of the details of global weather patterns into two neurons, but you can train one of them to tell you whether or not the climate is currently in an El Niño period, and that's a pretty solid basis for weather prediction. Or perhaps a weather prediction network would come up with some variable that had nothing to do with El Niño, and we'd learn something new.

There's no guarantee that this strategy will work for a given problem – some predictions are impossible to make

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 12 Nov 2026 at 09:24:50, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://www.cambridge.org/core/product/0258EEDF4C0E24D8F5A87BD4DB30E244

using so few variables. But training bottlenecked networks on interesting datasets could theoretically be a way to find useful, predictive concepts within them, if such concepts do exist.

One of the people who has worked most extensively with this idea, Raban Iten, is now CTO of an AI startup focused on nutrition. But when he was a physics graduate student in the late 2010s, he experimented with applying bottlenecked networks to problems in physics and achieved some promising results – so promising that he later wrote an entire book on using neural networks and other AI approaches for scientific discovery.[7] Like Ho, Iten used orbital mechanics as a testing ground, but he challenged his network with a trickier version of the problem of predicting planetary locations.[8] The data that he fed into the network were not the locations of each planet within the solar system but rather information about the positions of Mars and the Sun in the Earth's sky. This makes the problem much tougher: Earth and Mars both orbit around the Sun, so, when viewed from Earth, Mars follows a complicated, curlicued path. It's possible to use geometry to convert Earth-centric positions into Sun-centric ones in order to make predictions about a planet's trajectory, but that's a multistep process, and the math can get messy.

Presented with the problem of predicting how Mars will track across the Earth's sky, a symbolic regression system like the one that Udrescu and Tegmark proposed would be totally stumped, because carrying out that prediction requires more than one equation. A typical deep neural network, on the other hand, would certainly solve it, but it might not do so in a way that a human could understand. So

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 12 Nov 2026 at 09:24:50, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://www.cambridge.org/core/product/0258EEDF4C0E24D8F5A87BD4DB30E244

Iten, like Ho and Cranmer, constructed a constrained neural network. A few of the layers in his network had only two neurons each, which meant that all the information the network used to predict Mars's future location somehow had to flow through that two-neuron bottleneck.

By building his network with those tiny layers, Iten ensured that it would have to figure out how to represent all of the information that it needed to predict the locations of Mars and the Sun using only two variables. When Iten examined one of those layers, he found that its two neurons had learned to represent not Mars's and the Sun's positions from the perspective of the Earth – the information that the network was originally given – but rather Earth's and Mars's locations on their orbits around the Sun. The network had figured out the structure of our solar system, and the concept of heliocentric orbits, all on its own.

It's exciting to imagine that a relatively simple deep neural network could recapitulate Copernicus and independently discover the Sun's role in our solar system. But all of these orbital mechanics studies, while promising, are only proofs of concept. The scientists who run these studies know the solutions that the networks should find ahead of time, so it's possible that they are subtly pushing the networks in the right direction, even if entirely unintentionally. Moreover, orbital mechanics is child's play compared to many of the problems that stymie contemporary physicists. It still remains to be seen whether a deep neural network will ever uncover a new concept or formula that finds widespread use among physicists and meaningfully contributes to the field's progress.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 12 Nov 2026 at 09:24:50, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://www.cambridge.org/core/product/0258EEDF4C0E24D8F5A87BD4DB30E244

Nevertheless, working physicists are making extensive use of deep neural networks, to aid in tasks from running simulations to identifying interesting data points for further analysis. And though these networks do in some cases obviate the need for human expertise and understanding, in other cases they have helped researchers gain insight into problems far more timely than the nature of Mars's orbit.

Deep Neural Networks at the Large Hadron Collider

Physics is a huge field. Some physicists study objects so tiny they can't be directly observed; others take as their research subject the cosmos in its vast entirety. At both of these extremes, and in many of the subfields that lie in between, physicists are enthusiastically incorporating deep neural networks into their research. But it may be in particle physics that deep neural networks have found their warmest reception. After all, *WIRED*'s "End of Science" issue highlighted the LHC, the field's banner experimental facility, as an exemplar of the petabyte age. Particle physicists are constantly drowning in data, and deep neural networks help them stay afloat.

The goal of particle physics is to take the universe apart to reveal its most fundamental components. That might mean studying familiar objects, like protons, neutrons, and electrons, but often it involves focusing in on particles that are extremely unstable or otherwise challenging to observe. In order to generate those more interesting particles, scientists accelerate banal particles such as protons nearly up to

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 12 Nov 2026 at 09:24:50, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://www.cambridge.org/core/product/0258EEDF4C0E24D8F5A87BD4DB30E244

the speed of light and then smash them together. Getting particles to move that quickly requires enormous specialized facilities; the LHC, for example, shoots protons around and around a seventeen-mile-long, ring-shaped tunnel under the border between France and Switzerland to get them up to speed. By the time a pair of protons collides at the LHC, they have traveled hundreds of millions of miles – on the order of the distance from Earth to Jupiter.

At the LHC, the portions of the tunnel where protons collide are surrounded by layers of detectors that record the positions and energies of particles as they fly apart in the aftermath. Based on these detector hits, scientists can infer the identity and momentum of practically every product of a given proton–proton collision. But it's a difficult data processing task. Over the past decade physicists have worked to optimize that process using deep neural networks trained to perform various steps in the pipeline, from predicting which particles are electrons to tracking particles that take unusual paths through the detectors.

Unfortunately, interesting particles like the Higgs boson never make it to the detectors: They decay into other particles almost as soon as they are formed. It's these decay products that are directly measured, and knowing the identity of the decay products isn't enough to determine which particles gave rise to them. For example, the Standard Model – the overarching particle physics theory that predicted the existence of the Higgs boson decades before it was experimentally verified – says that the Higgs sometimes decays into a pair of photons. But other processes that happen at the LHC also produce two photons, and sometimes non-photon

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 12 Nov 2026 at 09:24:50, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://www.cambridge.org/core/product/0258EEDF4C0E24D8F5A87BD4DB30E244

objects are misclassified as photons by the LHC's data analysis algorithms. To make matters worse, the Higgs is only created in about one out of every billion proton collisions at the LHC, and only about one in every five hundred of those Higgs bosons will wind up as a photon pair. That means that particle physicists need to generate and analyze a prodigious volume of data in order to find anything interesting.

To detect the presence of the Higgs, physicists have to determine whether at least some of the photons they are observing come from Higgs decays rather than from other processes. In particle physics jargon, the target, Higgs-containing process is referred to as the signal, and the other processes are known as the background. Sometimes, physicists analyze detector data from tons of collisions all at once and look for statistical evidence of signal among all that background – that's how the Higgs boson was discovered back in 2012. At other times, they focus in on individual collision events. The goal in those cases is to analyze the data that the detectors record about each photon pair's energy and position to determine whether it was more likely to have been produced by a signal or background process.

Classifying events into categories based on data is a classic machine learning problem, and physicists have been using machine learning to solve it for a long time. But if you want to train a machine learning algorithm to tell apart signal and background events, you need to be able to identify those events in the first place. A computer vision model needs to know which images in its dataset are of cats, which are of dogs, and which are of Ferraris, so that it can learn to apply those labels to images it encounters in the future.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 12 Nov 2026 at 09:24:50, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://www.cambridge.org/core/product/0258EEDF4C0E24D8F5A87BD4DB30E244

Similarly, an algorithm that is supposed to distinguish detector data that indicate the presence of a Higgs boson from data produced by a background process needs plenty of labeled examples of both Higgs boson data and background data.

The problem is that scientists don't know ahead of time which data points are associated with the presence of the Higgs, because they can't observe the Higgs directly. So many of the classifiers used at the LHC are trained not on real data but instead on data generated by complex simulations. These simulations incorporate particle physicists' understanding of how particles form, interact, and decay following a high-energy proton collision, and because the collisions take place in a computer rather than in a real collider, physicists can observe every detail of each one. Classifiers use this labeled data to learn patterns to distinguish signal from background that, physicists hope, also hold for real data. (Traditional particle physics simulations can be expensive and time-consuming to run, and deep neural network-based simulations are often much faster – yet another way neural networks are being integrated into the experimental particle physics pipeline.)

Before 2012, physicists built their classifiers using a variety of machine learning techniques, including shallow neural networks, which only include one layer of neurons between the input and output layers. But shallow networks aren't nearly as effective at finding complex patterns in data as deep networks. To make the networks function effectively, physicists had to use their theoretical knowledge to transform the low-level detector data that they had about each event into high-level quantities that might be useful

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 12 Nov 2026 at 09:24:50, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://www.cambridge.org/core/product/0258EEDF4C0E24D8F5A87BD4DB30E244

for classification – for example, the total masses of various groups of observed particles. This is similar to how computer vision systems worked in the days before AlexNet: The machine learning approaches available at the time weren't powerful enough to learn from the low-level pixel data in images, so researchers first had to use human-designed algorithms to extract informative, high-level information from each picture.

By 2012, particle physicists already had plenty of machine learning experience, so they were ideally positioned to take advantage of the deep learning revolution. And the timing was fortuitous. The discovery of the Higgs boson, which many particle physicists had spent decades working toward, was announced in July 2012, so AlexNet's victory in the ImageNet competition that October came at a time when some researchers were considering how to spend the next phase of their careers.

Around that time, Daniel Whiteson, a particle physicist at UC Irvine who works extensively with machine learning tools, gave a talk at his university's computer science department where he spoke of his frustration at all the hand-holding demanded by shallow neural networks. In response, one of the computer scientists in the audience suggested he try out deep neural networks, which were still so new that Whiteson hadn't yet heard of them. So Whiteson set up a head-to-head battle between tried-and-true shallow networks and their cutting-edge cousins, in which he tested whether each network could learn to distinguish a Higgs boson signal from background, using either low-level data about the decay products or high-level mass calculations.[9]

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 12 Nov 2026 at 09:24:50, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://www.cambridge.org/core/product/0258EEDF4C0E24D8F5A87BD4DB30E244

As expected, the shallow network did well with the high-level mass data, and especially well if given both the high- and low-level data, but relatively poorly when it only had access to the low-level data. But the deep neural network was different – using only the low-level data, it blew the shallow network's best results out of the water. "You could just feed it the raw data and it would do really, really well," Whiteson says.

Interestingly, giving the deep network access to both the low-level data and the high-level data didn't improve its performance at all – either it had figured out how to extract the mass information from the low-level data itself, or it was using some other, equally effective means to make its predictions. And when it only had high-level data, it did markedly worse than it did with the low-level data. All the human expertise that went into extracting those high-level features from the LHC data was at best useless to the deep network, and at worst an active hindrance.

From one angle – the practical one – Whiteson's study suggested that deep neural networks could help physicists get more mileage out of expensive LHC experiments. The fact that the deep neural network performed better on the low-level data than on the high-level data indicated that physicists were inadvertently throwing out useful information when they processed their data before feeding it into machine learning algorithms. Deep neural networks provided a powerful, automated way of squeezing as much juice as possible out of raw LHC data.

From an intellectual angle, the study told a more complicated story. In some ways, it seemed like a boon for

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 12 Nov 2026 at 09:24:50, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://www.cambridge.org/core/product/0258EEDF4C0E24D8F5A87BD4DB30E244

theoretical progress in physics: By providing a clearer picture of what's going on inside the LHC, deep neural networks might help physicists test their theories more efficiently and provide inspiration for new theories. But the study also suggested that deep neural networks could render some types of theoretical work unnecessary. Physicists had worked hard for decades to understand the subatomic processes taking place inside particle colliders and leverage that understanding to better analyze the data that colliders produced. The end product of all that effort was the high-level data on which older classification techniques depended. Deep neural networks had no need for the high-level data, or for all the physical understanding that went into generating it.

As an experimental physicist who specializes in machine learning, Whiteson doesn't have to worry too much about the relevance of his own work. Nevertheless, he says, he isn't content to blithely build prediction tools and ignore the theoretical questions they hurdle over. "In the end, physics is about explaining the universe to humans, not to bots," he says. "We don't want AI [writing] papers for other AI that nobody can understand."

Making Networks Talk

When Whiteson introduced deep neural networks to the world of particle physics, not everyone was on board. One of the holdouts was Jesse Thaler. Thaler, a physicist at MIT, had spent much of his career devising theoretically grounded approaches for analyzing collider data – his most highly cited paper proposes a new high-level feature that can help

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 12 Nov 2026 at 09:24:50, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://www.cambridge.org/core/product/0258EEDF4C0E24D8F5A87BD4DB30E244

distinguish different types of jets, or narrow sprays of particles that are frequently produced in colliders. His job, more or less, was to come up with clever data processing algorithms grounded in his knowledge of physics. And now inanimate neural networks that knew nothing about the Standard Model were cooking up superior algorithms on their own. "I felt very threatened," Thaler says.

That's the context in which two graduate students, Patrick Komiske and Eric Metodiev, approached Thaler to show him some work they had done on applying deep neural networks to the problem of distinguishing jets produced by quarks, subatomic particles that make up larger particles such as protons and neutrons, from those produced by gluons, which bind quarks together into those larger particles (like glue). Coincidentally, Thaler had just finished working on a report with a group of physicists from around the world about why distinguishing quark jets from gluon jets is an ill-defined task. Like the Higgs, isolated quarks and gluons can never be observed directly, so their classifiers need to be trained on simulated data. But Thaler and the other physicists had just shown that the available simulators produced inconsistent jet data – so it was impossible to obtain a reliable set of training data, with some jets labeled "quark" and others labeled "gluon," from which a classifier could accurately learn to distinguish jets in the real world.[10] From Thaler's perspective, when Komiske and Metodiev trained a deep neural network on simulated data, the network hadn't really learned to tell apart quark and gluon jets at all.

To Thaler's surprise, Komiske and Metodiev agreed with him, and so the trio started looking for a better solution.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 12 Nov 2026 at 09:24:50, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://www.cambridge.org/core/product/0258EEDF4C0E24D8F5A87BD4DB30E244

Rather than abandoning machine learning entirely, they turned to a type of machine learning known as unsupervised learning. Many AI tools, such as AlphaFold, AlexNet, Whiteson's neural networks, and Metodiev and Komiske's earlier jet-classifying network, are products of supervised learning, in which algorithms are trained to associate data points with specific outputs. But in unsupervised learning, algorithms find patterns in datasets that don't come with predetermined labels or outputs.

In some cases, unsupervised learning can solve problems that might initially seem more amenable to a supervised approach. It's easy to train an algorithm to distinguish pictures of cats and pictures of dogs if you feed in a bunch of labeled examples – that's supervised learning. But it's also possible to teach an algorithm to separate a bunch of images of house pets into two separate categories, even if it's never told which images represent which species, because dogs (for the most part) have more in common with each other than they do with cats, and vice versa. That's unsupervised learning. The unsupervised algorithm might make some mistakes – it might get tripped up by chihuahuas, for example – but it will still bring to the surface important patterns, even without the explicit guidance that comes with supervised learning.

Unsupervised learning ended up being the perfect approach for the quark–gluon discrimination problem. By adapting a technique called topic modeling, which was originally used to identify the themes represented in a collection of documents in an unsupervised manner, the team devised a strategy for distinguishing quark and gluon jets that didn't

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 12 Nov 2026 at 09:24:50, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://www.cambridge.org/core/product/0258EEDF4C0E24D8F5A87BD4DB30E244

depend on dubious simulations.[11] In doing so, they resolved the dilemma that Thaler had helped to identify just about a year earlier. "The definition of the problem itself needed a machine learning language," Thaler says. "This was wild to me." He'd become a machine learning convert.

These days, Thaler serves as director of the National Science Foundation's Institute for Artificial Intelligence and Fundamental Interactions, which brings together researchers from AI and physics to accelerate both fields. Emblazoned across the Institute's website, in bold colors, is its unofficial motto: "Deep Learning (AI) + Deep Thinking (Physics) = Deeper Understanding." The phrase, Thaler told me, has its origins in his AI-skeptic era, when he was asked to comment on deep network-related research at a meeting and quipped that theoretical physicists should focus on deep thinking instead. Today it is, to all appearances, a mission statement for his career.

Though Thaler has become a strong advocate for deep learning in physics, he's not personally interested in the purely practical applications of deep neural networks. "My job, if I'm using machine learning, is to design machine learning algorithms where the outcome is the understanding that I'm seeking to gain," he says. So when, for another project, Thaler and his students trained a deep neural network to distinguish quark and gluon jets, they didn't use a vast, impenetrable network; instead, they intentionally designed a two-network system that could yield interpretable insights into what it had learned about the two kinds of jet.[12] The first network took in raw data about each individual particle in the jet and output several learned, high-level

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 12 Nov 2026 at 09:24:50, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://www.cambridge.org/core/product/0258EEDF4C0E24D8F5A87BD4DB30E244

features; the second network took those learned high-level features and predicted whether the jet had been produced by a quark or a gluon.

After training the networks, the team wanted to see if they could gain any new insights about particle jets from studying them. In order to make that task as straightforward as possible, they focused on the simplest version of their two-network system, which used only two high-level features to characterize each jet. One of these features, they found, focused exclusively on particles at the center of the jet, and the other focused only on particles at its edge. This was surprising: When humans had designed high-level features for quark/gluon classification based on their physics knowledge, they hadn't separated out particles at the center and the edge of the jets in this way. But when the team compared these new high-level features to human-designed high-level features – including one that Thaler himself had played a major role in formulating – the features uncovered by the neural networks appeared to do a better job of distinguishing quark and gluon jets.

For someone who had spent much of his career devising physics-informed ways to analyze collider data, extracting a novel approach from a neural network system was, as Thaler puts it, a "true revelation." By training a neural network-based system and then investigating what that system had learned, Thaler took a meaningful step toward greater physical understanding: not in a context of a toy problem like Newton's law of universal gravitation, but at the cutting edge of particle physics knowledge.

That said, Thaler's approach here is not a universal recipe for understanding what neural networks have learned

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 12 Nov 2026 at 09:24:50, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://www.cambridge.org/core/product/0258EEDF4C0E24D8F5A87BD4DB30E244

about the world, as he readily acknowledges. If Thaler, Komiske, and Metodiev had set out on this project with the goal of classifying quark and gluon jets as accurately as possible, it seems unlikely that they would have come up with their particular two-network scheme. And while some versions of their system did achieve classification accuracies that rivaled other top-tier approaches, the one from which they managed to glean new insights did not. The simplicity of that specific system meant that it was feasible to analyze it in detail, but it also made the system far worse than other, less transparent approaches at telling apart quark and gluon jets.

So if scientists want to gain knowledge from neural networks that have been designed for predictive accuracy rather than understandability, such as AlphaFold, they'll need a different strategy.

Lifting the Constraints

When large language models exploded into the public awareness in late 2022, they demonstrated, in dramatic fashion, the power of scale. The language models that power chatbots such as ChatGPT and Claude are gargantuan deep neural networks trained on vast volumes of data, and their impressive facility with language and problem-solving is partly attributable to their sheer size. These networks are sometimes called foundation models because they can be easily adapted for more specific language-related tasks, from answering medical questions to writing like someone from the seventeenth century, with a relatively small amount of additional training.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 12 Nov 2026 at 09:24:50, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://www.cambridge.org/core/product/0258EEDF4C0E24D8F5A87BD4DB30E244

Shirley Ho took the messages of that moment to heart. Within a year of ChatGPT's release, she, together with Miles Cranmer and other colleagues, founded an international research group devoted to developing foundation models for physics. So far, they have built and trained two huge foundation models: one that is specialized for handling various types of astronomical data, and another that can predict the movements of fluids and fluid-like systems, which is relevant for fields ranging from plasma physics to climate science. Their hope is that all sorts of different scientists will be able to adopt the models for their own purposes, without having to invest the time or money required to train a large-scale model from scratch.

Compared to the highly constrained model that Ho and Cranmer trained back in 2019 to predict the motion of objects under the influence of gravity, these foundation models are infinitely more useful. The gravity model had no real practical application: It was trained to perform a single task that physicists can easily complete without neural networks. The focus of that project was interpretability, not utility. These new foundation models, in contrast, have countless possible applications. The downside is that it's extraordinarily difficult to figure out how they work. "I don't think we have a very good way to interpret the models right now," Ho says.

When Ho and Cranmer built their gravity model, they intentionally chose a compact network whose structure closely matched the problem that they were trying to solve. Because of this correspondence, they were able to formulate well-grounded hypotheses about the roles that individual

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 12 Nov 2026 at 09:24:50, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://www.cambridge.org/core/product/0258EEDF4C0E24D8F5A87BD4DB30E244

layers would play in solving the problem, and those hypotheses turned out to be correct. But foundation models don't work like that. They abandon neuron-level correspondences in favor of scale and gain predictive power at the expense of interpretability.

The good news is that plenty of researchers – some of them very well resourced – are trying to figure out how large neural networks work, despite the challenges. Many of them work at major AI companies, such as Google DeepMind and Anthropic, where they probe the inner workings of proprietary large language models. For the companies, there's real value to understanding what's going on inside their models. If, for example, they were to find a set of neurons that was responsible for some undesirable chatbot behavior, such as giving users information about how to build a bioweapon, they could potentially curb that behavior by tweaking the activities of those neurons.

This subfield of AI research, which tries to link the behaviors of neural networks with the constituent parts of those networks, is called mechanistic interpretability. It's a challenging area. In simple, highly constrained neural networks, such as the network used by Ho and Cranmer to model gravity, individual neurons might play straightforward roles in solving the problem on which the network was trained. But that's typically not the case in large neural networks. A large language model doesn't have a specific neuron that responds when the network sees the word "cat" and another that responds when it sees the word "dog" – many neurons respond to each word, and some even respond to both. The meaning of a word is represented not by any

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 12 Nov 2026 at 09:24:50, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://www.cambridge.org/core/product/0258EEDF4C0E24D8F5A87BD4DB30E244

specific neuron's activity but rather by the overall pattern of activity in the network.

For that reason, figuring out which concepts a large neural network uses takes a bit of math. If each concept is represented by a pattern of activity across the whole network, those patterns need to be disentangled from one another somehow. One comparatively straightforward way to do this is to compare the network's activity pattern while it is representing a specific concept and while it is representing the concept's opposite – the difference between those two patterns will typically have a tight relationship to that concept. For example, a group of researchers led by Neel Nanda, who heads Google DeepMind's mechanistic interpretability team, compared the activity patterns in a large language model when it refused to comply with a user's request (say, to write a defamatory news article) with its activity patterns when it acceded to more benign requests.[13] When the team stimulated the neurons that were more active when the model refused, and suppressed the neurons that were more active when the model agreed, they could make it refuse to answer even the most anodyne queries, such as a question about the benefits of yoga. They seemed to have uncovered the population of neurons within the network that represented the concept of refusal.

The downside of this approach is that it only works if you have positive and negative examples of a specific concept ahead of time, so it leaves little room for the possibility that neural networks could help us uncover scientific concepts that haven't yet occurred to us. But other techniques fill that gap. One popular method involves training a separate

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 12 Nov 2026 at 09:24:50, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://www.cambridge.org/core/product/0258EEDF4C0E24D8F5A87BD4DB30E244

network called a sparse autoencoder to disentangle the concepts that are represented in a specific layer of a deep neural network. During training, the sparse autoencoder learns how to break down the original network's activation patterns into a set of more basic patterns that are thought to represent distinct concepts.

Because training a sparse autoencoder doesn't require the researcher to define any concepts ahead of time, it allows for the possibility that neural networks might uncover novel concepts on their own. Interestingly, when researchers do use this approach on large language models, they often find that the LLMs use concepts that seem to match human concepts. In one study conducted by researchers at Anthropic, for example, sparse autoencoders were used to disentangle activity patterns in one of the LLMs that underpins the company's Claude chatbot.[14] The experiments showed that one of these patterns became active whenever the LLM processed text related to the Golden Gate Bridge, and so the company briefly released a chatbot called "Golden Gate Claude" in which that neural pattern was artificially supercharged. The result was a chatbot monomaniacally fixated on San Francisco's most famous landmark. Ask a typical LLM about its physical form, and it will deny that it has one. Ask Golden Gate Claude the same question, and it responds, "I am the Golden Gate Bridge, a famous suspension bridge that spans the San Francisco Bay. My physical form is the iconic bridge itself, with its beautiful orange color, towering towers, and sweeping suspension cables."

Researchers are only just starting to apply these techniques to the problem of understanding what neural

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 12 Nov 2026 at 09:24:50, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://www.cambridge.org/core/product/0258EEDF4C0E24D8F5A87BD4DB30E244

networks trained on scientific data have learned. In late 2025, Miles Cranmer and his colleagues posted a paper that described their efforts to apply mechanistic interpretability techniques pioneered in large language models to Walrus, their fluid dynamics foundation model.[15] The results were promising: Using the difference approach, they found patterns of activity in the network that were strongly associated with familiar fluid dynamics concepts, such as vortices. When Walrus's vortex pattern was activated, the network would simulate vortices in fluids that had originally been flowing smoothly; when the pattern was suppressed, it would transform whirlpools into calm waters.

But knowing that a network appears to use some concepts familiar to human scientists is a far cry from fully understanding how it works. "There's still many steps to go before we will truly say, oh, we got it," says Ho. "It's very early days."

Perhaps new techniques that Cranmer and Ho could use to understand their foundation networks in more detail will emerge from the language model-focused mechanistic interpretability community. But that's far from guaranteed. Recently, Google DeepMind's mechanistic interpretability team – including Nanda, who is extremely influential in the field – posted an article advocating for what they termed "pragmatic interpretability."[16] Rather than striving to reverse-engineer neural networks, they argue, interpretability researchers should apply their skillset toward helping to ensure that AI progress has a positive impact on humanity – through, for example, finding the neural activity patterns that correspond to harmful LLM behaviors.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 12 Nov 2026 at 09:24:50, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://www.cambridge.org/core/product/0258EEDF4C0E24D8F5A87BD4DB30E244

Like many AI researchers, the authors of the article believe that AI could become quite powerful in the not-too-distant future, so aligning AI behavior with human desires and values is a pressing problem. That's partly why they endorse a pragmatic approach in which mechanistic interpretability researchers leverage their capabilities to make progress on practical AI problems as quickly as possible, instead of an intellectual approach in which researchers focus on the ultimate goal of understanding how AI systems work. It's entirely conceivable that the pragmatic approach could lead to progress in interpretability techniques – Nanda and his colleagues don't advocate for abandoning research into such techniques entirely, as those techniques could prove necessary to solving certain pragmatic problems. But such advancements seems less assured when mechanistic understanding itself is no longer the primary goal.

So it's not clear when, if ever, Ho and Cranmer will be able to figure out how their physics foundation models work. And even if they do, it's possible that answering that question won't help us understand the universe any better. When you train a highly constrained neural network, like Ho and Cranmer's gravity network, you can be relatively confident that the network will find the solution that you expect it to – its constraints leave it with limited options. But large, unconstrained neural networks have the capacity to find unanticipated patterns in the data, and those patterns might or might not facilitate human understanding. That's precisely why AlphaFold creator John Jumper doubts that figuring out what AlphaFold has learned will do much for the science of protein folding.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 12 Nov 2026 at 09:24:50, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://www.cambridge.org/core/product/0258EEDF4C0E24D8F5A87BD4DB30E244

If we want to determine whether neural networks have in fact learned anything interesting about the universe, we can't yet depend on mechanistic interpretability tools to help us out. "The big challenge about mechanistic interpretability is that it's very, very difficult to make conclusive claims about what a model understands if we don't fully understand the model," says Keyon Vafa, an AI researcher at Harvard. But that doesn't mean it's impossible to suss out what a network knows. Vafa and his collaborators have devised techniques that allow them to infer what a given neural network understands about the world without examining its innards – by, for example, observing how it behaves when it learns new tasks. If mechanistic interpretability is like neuroscience, Vafa's approach is like psychology.

In his research, Vafa has found that deep neural networks sometimes learn to make predictions in ways that bear little resemblance to how a human would solve the problem. One of his studies involved training a neural network to provide directions between pairs of locations in Manhattan, a task that a human would likely solve with a map.[17] The network did a great job of giving directions, but it didn't appear to have learned an accurate map of the city – it was like a child who parrots their parents' favorite phrases without understanding what they mean. In another study, Vafa trained a neural network to predict the positions of various planets as they orbited the Sun.[18] A physicist might solve this problem by using Newton's law of universal gravitation, but, when Vafa studied the model, he found that it didn't appear to have learned anything about Newtonian mechanics.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 12 Nov 2026 at 09:24:50, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://www.cambridge.org/core/product/0258EEDF4C0E24D8F5A87BD4DB30E244

In fairness, it is also possible for a human physicist to predict the position of the planets without drawing on Newton at all – they can always use Kepler's laws instead. But it's certainly the case that knowing Newton's law of universal gravitation allows someone to understand the movements of the planets at a deeper level. All together, Vafa says, his research calls into doubt the assumption that neural networks will automatically uncover the fundamental physical laws that undergird the data on which they are trained.

"I wish that [assumption] was true," he says. "I think it's unrealistic in many settings."

Superhuman Understanding

Figuring out what a trained neural network has learned about a physical system is a mighty problem, and solving it would be a major step forward. But even if we do manage to extract fundamental physical laws from neural networks, that knowledge still might not advance our scientific understanding. What happens if the methods and solutions that neural networks learn don't make any sense to us?

Even experts who are doggedly optimistic about the potential for neural networks to advance human understanding take this possibility seriously. "Do I think there will be machine learning algorithms that can make predictions that humans will not be able to understand? Absolutely," Thaler says. Krenn agrees. "In the probably near future, there will be computer solutions in physics which physicists cannot understand anymore," he tells me. "And the whole goal

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 12 Nov 2026 at 09:24:50, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://www.cambridge.org/core/product/0258EEDF4C0E24D8F5A87BD4DB30E244

of physics is to understand everything from first principles. So it's absolutely unacceptable."

Krenn is no fatalist about this scenario, which he calls the "dystopia of future science." We'll be okay, he says, as long as we develop tools that can help us break down those mind-boggling solutions into manageable chunks – tools that may very well be AI-based themselves. In principle, that might not be different from how humans teach each other. A student who has never taken an algebra class has almost no hope of understanding the solution to a linear equation the first time they see one, but with the help of a skilled math teacher, they can find their way to solving for x in no time.

But the teacher and student in this scenario have an advantage: Despite their different ages and levels of mathematical experience, they are both human. Deep neural networks, on the other hand, might solve problems in ways that are utterly alien to us. The philosopher Florian Boge, who holds a bachelor's degree in physics and is based at TU Dortmund in Germany, has argued that deep neural networks trained to make physics predictions could learn concepts that make little sense to humans and are of limited use for our scientific theorizing.[19] (He attributes this general line of argument to fellow philosopher Cameron Buckner.[20]) Instead of latching onto an easily interpretable concept like mass, a deep neural network could formulate its predictions using some byzantine combination of familiar quantities that is far too unwieldy for a human to use.

When I spoke with Boge, he described this scenario, or something like it, as "utterly plausible." To support this

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 12 Nov 2026 at 09:24:50, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://www.cambridge.org/core/product/0258EEDF4C0E24D8F5A87BD4DB30E244

view, he points to a phenomenon in computer vision known as an adversarial example. Back in 2014, a team of researchers discovered that it was possible to trick visual deep neural networks into mislabeling images by minutely tweaking the pixels in those images.[21] A skilled machine learning practitioner could take a picture of, say, the Eiffel Tower, and transform it so that a deep neural network would classify it as a cat, a coffee cup, or a car, even though a human would still easily identify each image as the Parisian landmark. Boge argues that such adversarial examples indicate that deep neural networks are picking up on image features that are undetectable or meaningless to humans. If that's true, fully explaining why the neural network classifies any given image as a cat or as the Eiffel Tower – in other words, determining "what the network knows" – would require referencing those counterintuitive features, and such an explanation would be extraordinarily difficult for humans to work with productively.

According to Boge, it's entirely conceivable that we could see features that are similarly meaningless or incomprehensible playing important roles in scientific neural networks. Even if we could extract the strange concepts learned by such a network, those concepts might do little or nothing to advance our scientific understanding.

Viewed from this angle, the problem of learning from deep neural networks resembles the conceit of Ted Chiang's novella "Story of Your Life" and its 2016 movie adaptation, *Arrival*.[22] Both the film and the novella center on a linguist, Louise Banks, who is recruited by the US government to figure out how to communicate with a group of aliens who have arrived to Earth. When the aliens show her

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 12 Nov 2026 at 09:24:50, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://www.cambridge.org/core/product/0258EEDF4C0E24D8F5A87BD4DB30E244

their writing system, she initially struggles to parse it – they weave together the symbols for individual words into complex, calligraphic images that don't bear any relationship to their speech. The way that the aliens compose these sentence-images is particularly confounding: When they start writing a sentence, they draw lines and shapes that participate in multiple word-symbols, as if they can somehow see how the entire sentence fits together before they have fully formulated it. Their approach defies the word-by-word logic of human language, and it's impossible for Banks to adequately reproduce their writing, at least at first. The alien writing system isn't understandable by human minds.

But over time, as she works with more and more of those images and continues to practice writing them, things start to shift: She begins to adopt a more alien perspective. The aliens, it turns out, don't progress through their lives linearly like humans do but rather experience all moments simultaneously, so they already know exactly what a sentence-image will look like before they put pen to paper. Banks never fully shifts to the alien mode of experience, but she does start to "remember" her future, and she also gains the ability to write fluently in the alien language. Though the alien writing system is initially impenetrable to her in fundamental ways, Banks works with the parts of it that she can access to the best of her ability, and it is through those first points of contact that she eventually comes to understand the entire system on its own terms. A radical expansion of her understanding begins with baby steps.

Perhaps we could make progress toward understanding the "alien perspectives" of sophisticated deep

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 12 Nov 2026 at 09:24:50, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://www.cambridge.org/core/product/0258EEDF4C0E24D8F5A87BD4DB30E244

neural networks in a similar way. One of the most visible advocates for this sort of approach is Been Kim, a Google DeepMind researcher who has been working in interpretability for over a decade. Back in 2018, Kim made waves at the Conference on Neural Information Processing Systems, the highlight of every neural network researcher's calendar, by sprinkling cold water on the field of interpretability research.[23] Through careful testing, Kim and her colleagues had shown that some popular techniques, which purported to identify the portions of an input that were relevant to the decisions made by a neural network, provided no insight into how the network actually worked. The study helped Kim establish her reputation as a creative and rigorous interpretability researcher, but she told me that it also sparked an "existential crisis" that she underwent a few years later. "I was like, 'Okay, I've been devoting my whole research life into interpretability and nothing is working, including everything that I did, and so I decided that maybe I should work on something else,'" she says.

But Kim wasn't able to tear herself away. Understanding what neural networks have learned, and ensuring that humans can take advantage of that understanding, was simply too important. "I kept coming back to this idea: But what about humans? Who's going to think about humans? Who's going to take the next step for humanity?" she says.

So Kim reoriented her research: Instead of trying to explain how a neural network works in its entirety, she wanted to find ways, however modest, of transferring at least some of the network's knowledge to humans. In defining

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 12 Nov 2026 at 09:24:50, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://www.cambridge.org/core/product/0258EEDF4C0E24D8F5A87BD4DB30E244

her new approach, she took inspiration from pedagogy. A good teacher doesn't blindly impose their own framework for thinking about a problem on their student; rather, they realistically appraise what the student knows and consider how best to move their knowledge forward. Learning occurs in a theoretical space known as the "zone of proximal development," where students work with ideas and skills that they can't yet manage independently but can successfully navigate with instructor guidance. Kim's goal now is to locate the zone of proximal development in which humans can indeed learn from AI systems, and to devise ways for us to get there. If the human doesn't end up grasping how the AI system works in its entirety, that's okay. The point is just to learn something.

To test this approach, Kim assigned the student role to some of humanity's most famed intellectual heavyweights: chess grandmasters.[24] Their teacher was AlphaZero, a chess-playing system designed by Google DeepMind that was, for a time, the best computer chess system in the world and remains far stronger than any human player. Rather than attempt to divine AlphaZero's entire chess-playing strategy and communicate it to the grandmasters – a task that seems nearly impossible, given how immature the field of interpretability research still is – Kim worked to discover a few concepts that the network used to identify promising moves. These concepts, which were encoded in the network's neural activity patterns, had to satisfy a few criteria. They had to be new for human players, so Kim focused on concepts that were more closely associated with AlphaZero chess games than with human chess games. They also had to be teachable,

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 12 Nov 2026 at 09:24:50, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://www.cambridge.org/core/product/0258EEDF4C0E24D8F5A87BD4DB30E244

so Kim tested whether they could be taught to a less sophisticated chess-playing network before showing them to human players. In this way, Kim targeted concepts that were neither too difficult nor too easy for the expert players – squarely in their zone of proximal development.

What Kim didn't do was check whether the concepts could be straightforwardly expressed in words. In fact, she didn't use words to teach the human players at all. Instead, she peered into AlphaZero as it played to identify sequences of moves in which the network appeared to be using each concept, and then she presented the players with the first board position in each sequence as a puzzle. After seeing a few examples of how AlphaZero solved the puzzles associated with a particular concept, the expert players seemed to get better at working with the concept themselves: Their puzzle performance improved. And, based on their commentary on the puzzles, they did seem to develop some explicit understanding of the ideas behind AlphaZero's moves – though it's difficult to evaluate how complete their understanding was.

Kim's work suggests that learning from a sophisticated deep neural network is possible, even if the network uses concepts that aren't straightforwardly expressible in human language. But a strategy that works for chess might not work in other areas. There's no obvious equivalent of chess puzzles in science, and even if a scientist did gain the ability to apply some convoluted neural network concept to concrete problems, it is unlikely they would be satisfied if they couldn't fully express how they were solving those problems to their peers.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 12 Nov 2026 at 09:24:50, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://www.cambridge.org/core/product/0258EEDF4C0E24D8F5A87BD4DB30E244

There's no guarantee, then, that researchers will be able to learn anything from complex scientific neural networks. But Kim is motivated to try. "I think there's a huge urgency," she says. "I have five-year-old and two-year-old kids. They're going to live in a different world. I want them to be an expert in something and have that satisfaction of understanding something so deeply and [knowing] they made the progress. I want them to keep that agency."

Two Paths for AI in Physics

If Kim succeeds beyond her wildest dreams and figures out how to teach humans what any deep neural network knows with minimal effort, the outlook for scientific understanding would seem to be bright. Deep neural networks would still in principle allow us to achieve prediction without understanding, but if extracting understanding from them is easy enough, the two objectives will still advance in near lockstep.

Of course, that's an enormous "if." As things stand now, physicists have indeed made meaningful progress toward extracting scientific insights from deep neural networks, but many of the researchers I spoke with see that line of work as distinct from more prediction-focused endeavors. "Curiosity-based research – which is what I do – curiosity-based science is all about human understanding," Thaler told me. "It's different if you're doing application-based science."

Whiteson, too, sees that divide, though he positions himself at least partly on the opposite side of it. Whereas Thaler is a theoretical physicist, Whiteson is an experimentalist; and though Whiteson has done some interpretability

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 12 Nov 2026 at 09:24:50, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://www.cambridge.org/core/product/0258EEDF4C0E24D8F5A87BD4DB30E244

studies, that line of work is not his primary focus. Whiteson prioritizes machine learning approaches that transform LHC data into informative predictions: Like molecular biologists who use AlphaFold predictions as starting points for their investigations of complex cellular pathways, he views deep neural networks as practical tools that can fold into larger scientific workflows. "I'm not always interested in the details of how [a network] works," he says. "I'm a tool user. I want to use the tool to ask questions about the universe."

One could imagine a fruitful collaboration in which Whiteson deploys a network to make scientifically useful predictions from LHC data and Thaler tries to understand the network's inner workings, and the two scientists have in fact worked together on some interpretability-focused research. But pursuing prediction and understanding simultaneously is much easier said than done, and neural networks that serve one goal might sacrifice the other. Researchers who aim for predictive accuracy have an extensive menu of neural network-based techniques at their disposal, and they have the freedom to choose whichever network gives them the best results. Researchers who prioritize interpretability, on the other hand, need to be deliberate when they design the networks that they use, in order to ensure that they will be able to extract insights from them later. As of now, many of the most impactful deep neural networks in science, such as AlphaFold, remain largely uninterpretable.

The idea that practical and intellectual goals might be opposed in machine learning is a common one, and it's been around for a while: Leo Breiman gave voice to it over

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 12 Nov 2026 at 09:24:50, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://www.cambridge.org/core/product/0258EEDF4C0E24D8F5A87BD4DB30E244

two decades ago when he wrote "Statistical Modeling: The Two Cultures." These days, machine learning researchers sometimes discuss the "accuracy–interpretability trade-off," or the idea that the approaches that yield the best predictions are also the least interpretable. Scientists debate whether or not this trade-off really holds true, and it certainly isn't applicable to every situation: Newton's law of universal gravitation is completely interpretable, and it does an exemplary job of predicting planetary trajectories when combined with his other laws of motion. But the trade-off has been observed in some settings.

And there's reason to think it applies to deep neural networks in physics. True, physicists have managed to build interpretable neural networks by transforming their theoretical knowledge into structural constraints on those networks. That's what Thaler and his team did with their two-network jet classifier, which incorporated their knowledge about the nature of jets; and that's also what Ho and her team did when they built their orbital mechanics model, which encoded key information ranging from the fact that planets exert forces on each other to some of Newton's laws of motion. But many machine learning experts believe that imposing such theory-driven constraints generally harms predictive performance in the long run.

This idea was most famously articulated by Richard Sutton, a computer scientist at the University of Alberta, in a widely discussed essay titled "The Bitter Lesson."[25] In just over a thousand words, Sutton efficiently marshals decades of experience in artificial intelligence to argue that data and scale generally beat cleverness and insight when it comes to

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 12 Nov 2026 at 09:24:50, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://www.cambridge.org/core/product/0258EEDF4C0E24D8F5A87BD4DB30E244

predictive accuracy. "Seeking an improvement that makes a difference in the shorter term, researchers seek to leverage their human knowledge of the domain, but the only thing that matters in the long run is the leveraging of computation," he writes. "These two need not run counter to each other, but in practice they tend to. Time spent on one is time not spent on the other."

There are clear signs of the bitter lesson even in interpretability-focused research: In Thaler's jet classifier paper, the less interpretable and less constrained networks performed significantly better than their more transparent brethren. Nevertheless, there are cases where the goals of prediction and interpretability do align. If scientists had access to infinite data about every phenomenon that interested them then they could blindly obey the bitter lesson and throw enormous deep neural networks at every problem in their path, confident that the networks would achieve predictive perfection. But data are never infinite; often, they are highly limited. In such cases, baking some key constraints of the problem directly into the neural network can help it do more with less – and, at the same time, make the network easier to interpret.

And sometimes imposing constraints on a network just makes sense. Tess Smidt, a machine learning researcher at MIT and a trained physicist, builds deep neural networks that respect the symmetries of physical problems. A neural network that predicts the force on a planet given the locations of all the other planets in its solar system, for example, should know that, if the entire solar system rotates 90 degrees, the force will rotate 90 degrees as well. It's possible

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 12 Nov 2026 at 09:24:50, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://www.cambridge.org/core/product/0258EEDF4C0E24D8F5A87BD4DB30E244

to learn that fact from lots and lots of data, but encoding it in the network from the get-go is often a more straightforward approach.

Compared to, say, embedding Newton's second law in the structure of a neural network, these symmetry constraints appear modest. But they too can lead to novel insights. "There's a lot of consequences to seemingly simple assumptions," Smidt says. "And so when you bake that into an algorithm, it actually shows you the consequences of those assumptions that may not be intuitive." In one study, Smidt demonstrated that her networks could uncover asymmetries in physical systems that might not immediately be apparent, and help scientists understand those asymmetries, in a surprising way: When the symmetry-respecting network fails to learn to make a prediction, that's a sign that there's an undetected asymmetry afoot.[26]

If scientists continue to find ways to incorporate theoretical knowledge into deep neural networks while also improving predictive accuracy, and if machine learning researchers continue to expand their arsenal of interpretability techniques, then it is possible that the networks that make the best predictions could also someday offer the most valuable scientific insights. Prediction wouldn't necessarily be dependent on understanding, but the two goals would still be closely linked. For now, though, that link is more indirect. The pursuit of practical AI tools does have benefits for researchers such as Thaler and Ho: They can appropriate any technological advancements made in the name of prediction for their own research. But that appropriation takes substantial time, effort, and talent. In a deep-learning world,

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 12 Nov 2026 at 09:24:50, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://www.cambridge.org/core/product/0258EEDF4C0E24D8F5A87BD4DB30E244

such an investment only makes sense if you value understanding for its own sake.

And physicists certainly do. As long as there are funders willing to support understanding-oriented research, scientists like Thaler and Ho will find new and creative ways to leverage neural networks to advance their understanding of the universe. With enough money, and a bit of luck, such research could mean that deep neural networks turn out to be a net positive for scientific understanding, at least within physics.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 12 Nov 2026 at 09:24:50, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://www.cambridge.org/core/product/0258EEDF4C0E24D8F5A87BD4DB30E244