4 Neuroscience

Letting Go of the Dream

Understanding is a powerful motivator. The UC Berkeley psychologist Alison Gopnik famously compared the feeling of achieving understanding to orgasm, and it'd not be an unreasonable analogy: There's a particular pleasure and satisfaction that comes when the pieces of an explanation suddenly fall into place in your mind.[1] The drive to understand is a potent urge, and it's a relief to have that urge fulfilled.

That's certainly true in physics, and it's also the case in neuroscience. But shifting the scope from the external universe to the tissue inside our skulls also changes the stakes of understanding. Explanations about how the brain works don't just satisfy our basic curiosity about the world around us: They also offer us insight into ourselves.

In principle, every single experience you have ever had, from that first shock of cold air as you emerged into the world to the placid contentment of an aimless day spent in a place you cherish, can be explained using neuroscience. But that's only in principle. In actuality, the human brain is so fiendishly complicated that we comprehend only a minuscule fraction of what is happening within it at any given moment. There's a tiny, semitransparent worm called a nematode that lives in compost bins and has only 302 neurons,

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 04:10:36, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://www.cambridge.org/core/product/CFCF20125BB1ED3E1D2D8A0A3FEED7A6

all of which scientists have mapped out in exhaustive detail – but we still don't fully grasp how that simplest of nervous systems works.

In recent years, deep neural networks have provided a way to transform ungainly neuroscientific datasets into accurate predictions, as they have done in molecular biology and physics. Such applications of neural networks come with all the understanding-related considerations that we've seen in earlier chapters. But neuroscientists also put deep neural networks to another purpose: They use them to construct computer models that mimic various brain areas. The neurons of a deep neural network can effectively stand in for biological neurons, and the network's layers can stand in for brain regions. Such neural networks are not mere tools for making predictions about the brain; they are themselves simulated brains, or parts of brains.

That has given neuroscientists a strong reason to want to understand how neural networks work. In physics, researchers strive to understand deep neural networks because they want to figure out what the networks know, in case that knowledge might inspire new theories or ideas. But in neuroscience, understanding a deep neural network means something altogether different. If the network is a good enough model for the brain, understanding how the network works could mean understanding the brain as well.

Even in the absence of deep neural networks, scientists would be discussing what it means to understand the brain – and whether we can ever hope to do so, given its complexity. Despite, or perhaps because of, those challenges, it is neuroscientists who have most fully faced up

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 04:10:36, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://www.cambridge.org/core/product/CFCF20125BB1ED3E1D2D8A0A3FEED7A6

to the implications of deep neural networks for scientific understanding. And rather than attempting to preserve the field's traditional notions of understanding, neuroscientists are redefining understanding in light of what, to them, now seems feasible. These are ambitious efforts driven by committed scientists who want to see their discipline bring as much benefit to the world as possible. They are also, at times, painful.

"Neuroscience is so deeply in crisis, and it's crazy that no one talks about it," says Konrad Kording, a University of Pennsylvania neuroscientist who has been closely involved in this transformation. "We are giving up on the old dream of neuroscience."

The Ideal of Neuroscientific Understanding

Narratives of modern neuroscience often begin with an unlikely character – not a scientist or a physician, but a railroad foreman from rural New England. One mid-September afternoon in 1848, with the Vermont air crisp and the leaves starting to change, Phineas Gage was working on building a new railway. When completed, the Rutland and Burlington Railroad would enable passengers to travel by rail from Boston or New York to the city of Burlington, Vermont, which sits within fifty miles of Canada. Gage was only twenty-five, but he oversaw the work of an entire gang of laborers, among whom he was reportedly a "great favorite."[2]

On that afternoon, Gage was packing explosive powder into a hole that had been drilled into some rock that stood in the railway's path, so that the rock could then be

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 04:10:36, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://www.cambridge.org/core/product/CFCF20125BB1ED3E1D2D8A0A3FEED7A6

blasted away. As he tamped down the explosive with an iron rod two-thirds as tall as he was, he accidentally hit the rod against the surrounding rock, sparking the explosive and propelling the rod directly into his face. The rod entered Gage's skull beneath his left eye, pushed the eye out of its socket, obliterated much of the left frontal lobe of his brain, and then shot out through the top of his skull.

Miraculously, Gage survived the blast – he may not even have lost consciousness. With the help of his men, he returned to the tavern where he was staying, and, though bleeding profusely, he at times managed to sit and even walk on his own. At the tavern, he was treated by the local physician, John Harlow, who probed the hole in his head with a finger, removed a barely attached piece of brain as well as some small pieces of skull, and, finally, dressed his wounds.

To Harlow's surprise, after less than a year of convalescence, Gage returned to a relatively normal life. He died a decade later.

Gage's near-complete physical recovery from such a grievous wound, decades before Joseph Lister introduced antiseptic techniques to medicine, makes his story worthy of retelling. But Gage is not remembered today for his triumph over his injury: He is invoked in neuroscience classrooms as an avatar for the essential role the frontal lobe plays in our behavior. According to Harlow, though Gage tried to return to his previous role once he had recovered, his employers would not rehire him because he was no longer the man his underlings had so highly regarded. Though physically fit for work, Gage had become inappropriate, impulsive, and irresponsible. "A child in his intellectual capacity and manifestations,"

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 04:10:36, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://www.cambridge.org/core/product/CFCF20125BB1ED3E1D2D8A0A3FEED7A6

Harlow wrote, "he has the animal passions of a strong man."[3] Gage eventually found work elsewhere, and he later spent years driving stagecoaches on a hundred-mile route through Chile – no doubt a mentally demanding job.

For early neuroscientists, Harlow's writings on Gage demonstrated that the brain's frontal lobes are specialized for functions such as self-control and planning – and, more broadly, that individual brain regions have their own particular functions. Armed with those conclusions, a neuroscientist could explain Gage's behavior change by noting that the portion of his brain that held the reins of his less considered impulses had effectively been destroyed. Gage's case was one of the first that allowed scientists to map specific psychological capacities to brain areas, and, in doing so, begin to understand how the brain gives rise to behaviors.

As the field of neuroscience advanced far beyond its dependence on case reports of medical oddities, the explanatory strategy of linking physical brain structures to abstract functions nevertheless remained remarkably consistent. The Washington University philosopher Carl Craver, in a 2007 book titled, simply, *Explaining the Brain*, argues that all explanations in neuroscience are mechanistic; they involve taking a brain system of interest (perhaps the entire brain), breaking it down into physical parts, determining the jobs of all of those individual parts, and then figuring out how those parts work together to accomplish some greater goal.[4] The explanation of Gage's behavior is a minimal example of a mechanistic explanation – it specifies only one brain part, the frontal lobe – but it fits the overall pattern of mapping brain substructures to specific functions.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 04:10:36, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://www.cambridge.org/core/product/CFCF20125BB1ED3E1D2D8A0A3FEED7A6

A more typical example of mechanistic explanation in neuroscience came in the late 1950s, when neuroscientists David Hubel and Torsten Wiesel decided to start sticking electrodes in the brains of cats. By then, scientists had sussed out the roles of various brain regions in far more detail than Gage's case had enabled, but almost nothing was known about the specific jobs carried out by the individual neurons that made up those regions – largely because studying neurons was extremely difficult.

Neurons convey information using near-instantaneous bursts of electrical activity known as action potentials. In scientific parlance, a neuron that is undergoing action potentials is said to be firing. It's possible to record the action potentials of a neuron in a living animal by inserting a metal electrode directly into its brain tissue, but, in the 1950s, keeping an electrode in close enough proximity to a minuscule neuron, as the brain regularly pulsed with blood, was challenging. Hubel, however, had managed to design a remarkably stable electrode, and this new tool allowed him and Wiesel to examine the brain at the level of individual neurons.

Like many of their contemporaries, and many neuroscientists since, Hubel and Wiesel decided to focus their research on vision. Scientists had learned from experiments on brain-damaged dogs in the late nineteenth century that the brain's occipital lobe, located at the back of the head, is responsible for the sense of sight. Hubel and Wiesel wanted to dig deeper and find out exactly how that region worked. They were inspired by the work of their mentor Stephen Kuffler, who had demonstrated that individual neurons in the retina – which is often considered to be part of the

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 04:10:36, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://www.cambridge.org/core/product/CFCF20125BB1ED3E1D2D8A0A3FEED7A6

brain – tend to fire in response to dots of light in specific locations.[5] Some neurons prefer dots of light at the very center of the area that an animal can see; others might prefer dots at the upper left corner, or at the bottom right, or in any other place.

In light of Kuffler's work, Hubel and Wiesel began their study of the occipital lobe by showing cats spots of light at various locations. Today, scientists would use a computer to generate all of those stimuli, but Hubel and Wiesel had to get creative: To make the bright spots they needed, they shone light through a piece of metal with a small hole. In order to easily monitor the neurons that they were recording, they connected their electrode to an audio device that would emit a click whenever a neuron fired. Shining those spots of light at a cat's eye would have elicited robust bursts of noise if they had been recording from Kuffler's retinal cells, but Hubel and Wiesel failed to find any cells in the occipital lobe that would consistently fire in response to dot stimuli.

Their breakthrough came by chance. One day, as they passed a glass slide in front of the light source, they heard an unmistakable series of clicks. Puzzled, they struggled for a while to recreate the response until they realized that the neuron was reacting to the dark line produced by the edge of their slide. The neurons didn't care about the dots: They were interested in bars.

Hubel and Wiesel eventually found numerous cells that responded to bars. Each one had a preference for a bar located at a specific location and oriented at a specific angle.[6] This alone was a groundbreaking discovery, and it afforded

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 04:10:36, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://www.cambridge.org/core/product/CFCF20125BB1ED3E1D2D8A0A3FEED7A6

scientists unprecedented insight into how vision works. Outside the artificial conditions of the laboratory, these bar-responsive cells, which Hubel and Wiesel named simple cells, tend to fire where a dark object meets a light background, or vice versa – in other words, they detect edges. This edge detection ability is key to the visual system's ability to isolate and identify individual objects.

On its own, Hubel and Wiesel's description of simple cells could be considered a mechanistic explanation of how one small component of vision works, much as the link between the frontal lobe and self-control provides a mechanistic explanation of Gage's behavior change. But Hubel and Wiesel went farther: They built theories of how different groups of specialized neurons worked together to enable us to see. In 1962, near the end of a nearly fifty-page paper detailing the results of countless field-defining experiments, they proposed a mechanism for how the dot cells found early in the visual pathway might enable simple cells to respond to their characteristic bar stimuli.[7] A bar of light, they proposed, could be thought of as a set of overlapping dots of light, all arrayed in a straight line. If all the cells that responded to those dots sent signals to a single cell, and that cell only fired when it received signals from all those cells at once, then that cell would be a simple cell: It would fire if, and only if, the animal saw a bar of light.

This elegant story about neurons with clearly defined jobs instantiating the sense of sight is so influential in neuroscience that several of the researchers I spoke with for this chapter mentioned Hubel and Wiesel in our conversations, all without my prompting. More than sixty years later, the

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 04:10:36, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://www.cambridge.org/core/product/CFCF20125BB1ED3E1D2D8A0A3FEED7A6

pair's celebrated experiments and theories still represent the pinnacle of mechanistic explanation in neuroscience, and countless successors have held their achievements in mind when questing their own electrodes through the brains of cats, rats, birds, and primates. That's what Kording meant when he spoke of the "old dream of neuroscience": the idea that research could proceed through the brain indefinitely in the manner of Hubel and Wiesel, sussing out the functional roles of neurons one by one.

Data Overwhelm

For the old dream of neuroscience to work, neuroscientists would need more data – a lot more. Hubel and Wiesel had described only a minuscule fraction of the cells in the occipital lobe, much less the entire brain; 1950s-style electrodes, which can record from at most a few neurons at a time, were not up to the task of classifying and characterizing the billion cells in a cat's brain, or the tens of billions in ours. Over the past decade, however, experimental techniques in neuroscience have advanced considerably, and the field has fully entered its big data era. One of the most significant advancements came in 2017, when researchers introduced a new type of electrode, known as a Neuropixels probe, that can detect action potentials from hundreds of neurons at once.[8] Using Neuropixels and other tools, Seattle's Allen Institute for Brain Science has amassed enormous datasets covering both the mouse brain and the human brain, and they have made those data freely available to scientists. Other tools, like electron microscopy, have allowed researchers to probe

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 04:10:36, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://www.cambridge.org/core/product/CFCF20125BB1ED3E1D2D8A0A3FEED7A6

the structure of the brain with unprecedented precision: In 2024, an international team of scientists managed to map out all of the 140,000 neurons in the brain of a fruit fly, as well as the 50 million connections among them.[9]

Each of these datasets stands as an enormous scientific accomplishment in its own right. Acquiring, compiling, and releasing them demanded prodigious effort and ingenuity, and the scientists who worked on them have done a great service to their field. So far, however, large-scale datasets have not unlocked a new era of neuroscientific insight. The gap between the volume of brain data available and our understanding of how the brain actually works is perhaps bigger than it has ever been.

For Eva Dyer, a University of Pennsylvania professor who works at the intersection of machine learning and neuroscience, this glut of data, and the relative lack of understanding that came with it, functioned as something of a wake-up call. "All of a sudden, we were like, 'Wait, okay, we have all this data, it's spanning all these different regions, like thousands of different neurons, and we were saying that we wanted that all along,'" she says. "But then you have all this data and it's like, 'Oh, yeah, how do we actually make sense of it?'"

Shortly before Neuropixels emerged on the scene, Kording and his colleague, Eric Jonas, decided to directly attack the question of whether data alone could solve the brain. As they saw it, many scientists believed that a lack of data was the primary obstacle standing in the way of neuroscientific progress. To test that idea, they used neuroscientific techniques to study a system from which they could

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 04:10:36, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://www.cambridge.org/core/product/CFCF20125BB1ED3E1D2D8A0A3FEED7A6

collect ostensibly unlimited data: a simulated computer chip, on which they could play classic arcade games.[10] Whereas Hubel and Wiesel recorded from neurons in cat brains as the animals looked at blobs of light, Jonas and Kording recorded from transistors on a chip as the chip ran Donkey Kong. They tried out other neuroscience tricks as well, from removing individual transistors and seeing if the chip could still run games properly to analyzing the connections among the transistors with clever math. But none of these tactics brought them anywhere close to understanding how the chip worked.

Jonas and Kording didn't show that it's impossible to understand a computer chip using neuroscience-inspired techniques – their work wasn't exhaustive. But the study suggested that data poverty was far from the only challenge facing neuroscientists. If established neuroscientific techniques couldn't even conquer a 3,500-transistor computer chip, they were unlikely to solve the human brain.

There are at least two major reasons why scaling up brain data has not concomitantly scaled up neuroscientific understanding. The first, and simpler, reason is that humans are easily overwhelmed by large volumes of information. Hubel and Wiesel were so successful in part because they had the capacity to study each neuron from which they recorded in great detail. But it is difficult to imagine a scientist today noticing a strange blip from one of the hundred neurons that their Neuropixel probe detects and then spending the rest of the day struggling to determine the origin of that response, as Hubel and Wiesel did when they discovered the cell that responded to the edge of the

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 04:10:36, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://www.cambridge.org/core/product/CFCF20125BB1ED3E1D2D8A0A3FEED7A6

glass slide. "That whole handcrafted approach only works because you were dealing with very human-scale recordings," says Mazviita Chirimuuta, a philosopher of neuroscience at the University of Edinburgh who trained in visual neuroscience.

Even if a neuroscientist did manage to successfully characterize each of the thousands of neurons they observed, they might not be able to make much sense of those characterizations. The mechanism that Hubel and Wiesel proposed to explain why simple cells respond to bars of light – namely, that they receive inputs from center-surround cells that respond to different spots of light along that bar – is effective because it is compact enough for a person to hold in their mind and apply to new situations. But we cognitively limited humans can only handle so much information at once. A mechanism containing hundreds or thousands of different types of neurons might not help anyone understand anything.

The second reason is that most neurons are far more complex than the cells that Hubel and Wiesel studied. The mechanistic approach only works when it is possible to break a system down into individual components that have human-understandable functions and can be analyzed in isolation. In Hubel and Wiesel's case, those components were occipital lobe neurons, and those functions were defined by the specific, and straightforward, patterns of light to which those neurons responded. But in much of the brain, including in the areas that Hubel and Wiesel studied, neurons work collaboratively. Contrary to the myths of "grandmother cells" and "Jennifer Aniston neurons," your brain doesn't have a

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 04:10:36, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://www.cambridge.org/core/product/CFCF20125BB1ED3E1D2D8A0A3FEED7A6

specific neuron for each person that you know or recognize; rather, when you see or think of a particular person, a whole group of neurons fires to represent that information. Each of those neurons also helps represent many other people, much like each letter of the alphabet appears in countless different words. It's the overall pattern of letters that allows you to identify a specific word, and the overall pattern of firing in a population of neurons that conveys information about a particular person or idea.

This means that it's often impossible to figure out what a particular neuron does by looking at it in isolation, à la Hubel and Wiesel – its responses might be too complicated to make sense to us, or too entangled with the activities of other neurons. Blake Richards, now a professor of neurology and computer science at McGill University, came to this realization early in his neuroscientific career. While spending a few months in a visual neuroscience lab at the start of graduate school, Richards sat in on some experiments that involved recording from neurons in the visual cortex of a macaque. As the researcher inserted an electrode into the macaque's brain, they checked whether each neuron they encountered was orientation selective – that is, whether it fired more in response to bars of light at a specific angle, like the cells that Hubel and Wiesel identified. If the neuron didn't show that kind of response, the researcher moved on. Only the orientation-selective cells were worthy of further examination.

Richards was stunned that neuroscientists would completely ignore so much of their object of study. "I said to my supervisor at the time, 'But you're not studying any of

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 04:10:36, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://www.cambridge.org/core/product/CFCF20125BB1ED3E1D2D8A0A3FEED7A6

the other neurons!'" Richards says. "And he was like, 'Yeah, well, that's the way of it. We don't have any way to understand what those other neurons are doing.'" The researchers were attacking the problem of how animals see in the manner of the proverbial drunk man who only searches for his keys in the small circle of illumination cast by a streetlight, though he knows he dropped them elsewhere.

Soon after, Richards encountered a paper estimating that, despite decades of effort, scientists only understood about 15 percent of what was going on in relatively simple parts of the visual system – partly because of their near-exclusive focus on neurons that responded to stimuli in predictable and interpretable ways.[11] He began to think, he says, that the handcrafted approach to understanding the brain was doomed to fail. The brain was chock-full of neurons whose responses could not be straightforwardly described or effectively analyzed in isolation, even in the very region that Hubel and Wiesel had ostensibly understood so well. Neuroscience needed new tools that faced up to the brain's complexity.

Modeling the Brain

A lot of the energy in neuroscience these days, particularly in the field's mathier corners, is aimed at parsing and analyzing the big, convoluted datasets produced by tools such as Neuropixels. One approach, called dimensionality reduction, tries to identify a short list of factors that can capture most of the interesting information contained in recordings of hundreds or thousands of neurons – like extracting the various tropes found across a collection of romance novels,

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 04:10:36, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://www.cambridge.org/core/product/CFCF20125BB1ED3E1D2D8A0A3FEED7A6

rather than focusing in on every individual word of every text. Another approach, known as dynamical systems theory, focuses on how the activity of a population of neurons changes over time and tries to capture those patterns in a simple mathematical form.

And then, of course, there's machine learning. In big neuroscientific datasets, each chunk of brain activity data is typically accompanied by some information about what the animal was experiencing or doing at the time: the specific visual stimulus it was viewing, for example, or its specific location in a maze. It's natural to ask, then, whether it's possible to reconstruct those actions or experiences based only on the neural data. The answer is often yes. Treating an animal's brain data as the input and its location as the output, you might be able to train a machine learning algorithm to predict where an animal is in space based only on which of its neurons are currently firing.

Such "decoding" tools are impressive and uncanny in equal measure. Back in 2011, before AlexNet, a team from Berkeley ran noninvasive brain scans on a few people while they watched movies and then trained a basic machine learning method to reconstruct those movies from their brain data.[12] The reconstructed movies are far from perfect – people and objects look ill-defined and blob-like, and the colors are all wrong – but the general shapes and motion in each shot are remarkably spot-on. In partially extracting the movies from their subjects' brain data, the scientists were, in effect, reading their minds.

With the introduction of deep neural networks, video reconstruction from brain scans got far more accurate.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 04:10:36, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://www.cambridge.org/core/product/CFCF20125BB1ED3E1D2D8A0A3FEED7A6

So did predicting speech from neural data. For almost two decades, neuroscientists have been working to restore speech to people who are completely paralyzed. These people receive brain implants that record the activity of individual neurons in the brain area responsible for speech; the task of the researchers is to translate those neural firings into words. Early attempts were inaccurate and laggy, but more recent neural network-powered methods can produce speech from neural data almost instantly and even decode the speaker's intended intonation.[13]

"It's magical," Kording says. "We are literally giving back speech to people who lost it. That is a biblical capability – it's literally in the book."

At a high level, the deep neural networks used for decoding are comparable to networks such as AlphaFold and AlexNet. They function as tools, like microscopes or calipers: The predictions that they enable might be practically or scientifically valuable, but the tools themselves are not primary objects of study. But in neuroscience, unlike in molecular biology and in physics, deep neural networks aren't primarily used as tools. As impactful as decoding research may be, to neuroscientists, deep neural networks are most widely recognizable as models of how the brain works.

Throughout their history, neural networks have had a close relationship with neuroscience. Frank Rosenblatt, who proposed the first multilayered neural network in 1958, saw his networks not as tools from learning from data but rather as models that might help scientists understand how the brain processes information.[14] These days, some deep

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 04:10:36, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://www.cambridge.org/core/product/CFCF20125BB1ED3E1D2D8A0A3FEED7A6

neural networks resemble specific brain circuits: The class of neural networks that includes AlexNet, for example, was directly inspired by Hubel and Wiesel's research. All of these connections make deep neural networks particularly apt devices for building models of the brain, or parts of it, in a computer. "The whole brain is at some level a deep network," says Jim DiCarlo, an MIT neuroscientist who, like Hubel and Wiesel, studies how vision works in the brain.

In the pre-AlexNet days, when the rest of the world was paying deep neural networks little mind, DiCarlo recognized their potential utility for neuroscience. He says that he'd sometimes find himself alone at computer science conferences with Yann LeCun, now a world-famous deep neural network pioneer, because few other attendees saw promise in neural networks. Then AlexNet won the ImageNet challenge, and LeCun, who had laid down much of the groundwork on which AlexNet was built, became a celebrity. DiCarlo and his laboratory quickly started working with AlexNet-like networks, and in 2014 they released a blockbuster paper that firmly established the utility of deep neural networks for visual neuroscience.[15]

The basic idea of the paper was simple. DiCarlo had at his disposal two very different tools that could identify objects in images – trained deep neural networks, and the macaques that served as the laboratory's test subjects – and he wanted to compare them in order to find out whether they solved the problem of object identification in a similar way. Because the deep neural networks that DiCarlo was working with were originally inspired by Hubel and Wiesel's research, they already resembled the brain at a

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 04:10:36, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://www.cambridge.org/core/product/CFCF20125BB1ED3E1D2D8A0A3FEED7A6

high level: Just as visual information flows sequentially through the layers in a deep neural network, it flows through a hierarchical series of regions in the macaque's brain. So DiCarlo paired up network layers with brain regions and looked for similarities. When he and his team compared the neurons in the final layer of the neural network to those found in the macaque's inferior temporal cortex, a region that is highly associated with object recognition, they found a surprisingly close match. And when they compared the second-to-last layer in the network with the brain region that immediately precedes the inferior temporal cortex in the visual pathway, they observed the same correspondence.

Crucially, the network hadn't been explicitly trained to function like the brain; instead, as it learned to complete the same object identification task that the macaques performed, its neurons spontaneously started to resemble macaque neurons in terms of how they responded to various images. While working to capture an input–output relationship between images and object labels, the network seemed to have learned something about how the macaque brain worked.

In the years since DiCarlo undertook his seminal study, other neuroscientists have shown that deep neural networks can effectively model neural circuits that undergird hearing, smell, touch, and spatial navigation. There's also some intriguing recent work suggesting that large language models might work similarly to the brain's language-processing networks. But visual neuroscience is the subfield that has been most dramatically transformed by the rise of deep neural networks. These days, according to DiCarlo,

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 04:10:36, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://www.cambridge.org/core/product/CFCF20125BB1ED3E1D2D8A0A3FEED7A6

it's unusual to study how vision works in the brains of adult mammals without using neural networks.

And many visual neuroscientists have embraced them wholeheartedly. Among them is Grace Lindsay, a New York University neuroscientist who studies how attention works in the brain. She encountered DiCarlo's work on neural networks while studying much simpler models of the visual system in graduate school. Those models were easy to understand, but Lindsay had grown frustrated at their small scale, as they only captured a tiny part of the brain's visual processing pathways. "I always wanted to know, okay, but then what happens next?" she says. "Where does this information go after this circuit has processed it?"

When Lindsay read DiCarlo's paper, she saw that the networks he and his team had trained didn't just mimic one small part of the macaque's visual system: They modeled how visual information might flow from the retina to the brain's complex hierarchy of visual regions. "I was like, 'Oh, so this is what I need,'" she says. "This is a model that I can work with to really connect how neural activity changes cause changes in perception and behavior."

Lindsay is now a professor, and she specializes in neural networks. She is part of a large and growing community of neuroscientists who, impressed by the ability of neural networks to model the brain with unprecedented scale and accuracy, have centered their research around them. Another member of that community is Nikolaus Kriegeskorte, a Columbia neuroscientist who, like DiCarlo, published an influential paper in 2014 that demonstrated the uncanny resemblance between macaque brains and

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 04:10:36, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://www.cambridge.org/core/product/CFCF20125BB1ED3E1D2D8A0A3FEED7A6

deep neural networks.[16] Kriegeskorte credits deep neural networks with accelerating the development of a new neuroscientific subfield known as cognitive computational neuroscience, which studies both brain circuits and cognitive processes using detailed, end-to-end models of vision and other mental capabilities. That field has had an annual conference since 2017; in 2024, five out of six invited keynote lectures at the event focused on deep neural networks.

Before deep neural networks, scientists had to decide between studying the brain from a bottom-up perspective – by, for example, building biologically inspired models of small neural circuits – or taking a more top-down approach that involved observing how humans actually think and behave. But deep neural networks allow scientists to do both at the same time: The networks incorporate specific, neuron-level details, but they can also perform sophisticated tasks such as image identification and language production. They make it possible to bridge the neural and the psychological.

"Without them," Kriegeskorte says, "we're totally lost."

Redefining Understanding

But not everyone welcomed deep neural networks with open arms. "A lot of neuroscientists were pretty resistant to using these kinds of models," Lindsay says. When she started giving talks on her neural network research, audience members would consistently ask her to defend the choice to use deep neural networks as models of the brain. Eventually,

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 04:10:36, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://www.cambridge.org/core/product/CFCF20125BB1ED3E1D2D8A0A3FEED7A6

she grew so tired of answering the same questions all the time that she added a set of introductory slides to her talks, in order to preempt them.

Some of Lindsay's interlocutors were skeptical that deep neural networks were as effective and useful for modeling the visual system as she claimed. The models Lindsay had previously worked with, which had been common in visual neuroscience, were carefully handcrafted based on observations of the brain's anatomy and theories about its functioning. But although the overall structure of deep neural networks is inspired by the brain, their specific details are automatically learned from data, not manually designed based on neuroscientific principles. Could they really be considered brain models, if they hadn't been deliberately built to match the brain's anatomy?

For that particular point, Lindsay had a clear counterargument. Compared to earlier models of the visual system, deep neural networks appear to do a far better job of capturing what is going on in the brain when an animal sees an image. Unlike those standard models, neural networks can respond to and process entire images, like an animal does; and as they do so, they match the activity patterns in the animal's brain more closely than other models do. Though the networks are not carefully constructed piece by piece to mimic the mammalian visual system, they end up resembling it regardless. That was more than enough for Lindsay's purposes.

There's another common concern about deep neural networks in neuroscience that is trickier to defuse. Many neuroscientists see the goal of their work as understanding

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 04:10:36, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://www.cambridge.org/core/product/CFCF20125BB1ED3E1D2D8A0A3FEED7A6

the brain: answering how and why questions about how our perceptions, experiences, and behaviors are instantiated in the three pounds of gelatinous flesh between our ears. Given that we struggle to answer such questions about how deep neural networks function, some scientists doubt that these networks can advance neuroscience. By using deep neural networks to study the brain, researchers are simply substituting a silicon black box for a biological one – and entirely failing to advance neuroscientific understanding.

Though Lindsay disagrees that studying deep neural networks means giving up on understanding the brain entirely, she does see where these concerns come from. "When I learned about these models in the middle of my PhD, [it] kind of shook up my thinking about what an appropriate research question is," she says. "The idea that we're going to be able to say, 'This is the role that this neuron plays in object recognition,' I just think that shouldn't be the goal anymore."

As Lindsay sees it, deep neural networks aren't compatible with a Hubel and Wiesel-like approach to parsing the brain's visual circuitry. "And that's okay," she says. "We don't need to understand the brain that way."

Lindsay is far from the only neuroscientist who's thinking seriously about understanding. The rapid rise of deep neural networks has forced many of her peers to confront questions previously restricted to the book-lined offices and well-worn armchairs of philosophy departments. What does it mean to understand the brain anyway? What sorts of understanding, if any, can deep neural networks offer? And are we content with those kinds of understanding?

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 04:10:36, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://www.cambridge.org/core/product/CFCF20125BB1ED3E1D2D8A0A3FEED7A6

Like Lindsay, DiCarlo was met with skeptical responses when he presented his early work with deep neural networks. "People would raise their hands and say, 'That's a little disappointing – we don't understand that network any more than we understand the visual system you're recording from,'" he says. Those comments got DiCarlo thinking about what deep neural networks might have to offer, even if they couldn't provide human-interpretable, mechanistic explanations of how vision worked.

DiCarlo had already demonstrated that deep neural network models can furnish remarkably accurate predictions about how a specific neuron in a macaque's brain will respond to a particular image. He realized that such robust predictions brought with them another capability: control. If you have access to a deep neural network that effectively predicts how a neuron in a macaque's brain reacts to any image, you can construct new images that activate that neuron far more strongly than anything else the macaque has previously seen, using the network as a guide. It's a similar concept to using protein folding networks such as AlphaFold to design a protein that take on a desired shape.

When DiCarlo and his team applied this technique, they obtained images that didn't look like anything a macaque was likely to see in nature.[17] They resembled abstract fabric patterns, or hyper-zoomed-in pictures of a reptile's skin. But the technique worked: Many of the images activated neurons in the macaque's brain more strongly, and more specifically, than any natural image would.

In the article in which DiCarlo and his team presented their work, they described the experiment as "a

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 04:10:36, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://www.cambridge.org/core/product/CFCF20125BB1ED3E1D2D8A0A3FEED7A6

practical test of useful, causal 'understanding.'"[18] The implication was that the ability to control neurons is equivalent to the ability to understand them. But many philosophers would argue that those concepts are distinct. According to them, DiCarlo doesn't necessarily understand where the generated images came from in any depth; he's trusting the neural network's grasp on the situation, much like someone reading off celestial predictions from Kepler's ephemerides trusts that Kepler himself understands the solar system.

DiCarlo acknowledges that some people find this notion of understanding unsatisfying. "We asked the model in the computer how to make the neurons go into [some] state, and it provided an answer, and the humans would really like to be the ones providing the answer," he says. But he contends that this perspective loses sight of the ultimate goals of neuroscience. "We're not building these models to make scientists happy," he says. "We're supposed to build them to make the world better."

There are other ideas floating around about what it might look like to understand the brain using deep neural networks. One influential paper from 2019, entitled "A Deep Learning Framework for Neuroscience," proposed that neuroscience deprioritize discovering mechanisms in favor of forms of understanding that might be more easily gleaned from neural networks.[19] It arose out of discussions at a workshop that Richards, Kording, and Lindsay all attended; Richards and Kording both played major roles in writing it.

The paper starts by specifying some ways in which it is possible to understand deep neural networks. Although

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 04:10:36, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://www.cambridge.org/core/product/CFCF20125BB1ED3E1D2D8A0A3FEED7A6

scientists have little comprehension of the mechanisms that allow a deep neural network to, say, identify a dog in a photo, they can answer some questions about how and why the network is so good at object identification. They know the network's overall structure – the number of layers it has and the ways in which the layers are connected to each other – and they know that networks with other structures aren't as good at that task. They also know a lot about the task on which the network was trained, from the specific images the network was asked to label to the way it received feedback when it guessed incorrectly and the mathematical rules that dictated how it learned during that training process. All of these facts are good answers to questions about why the network is so good at its job. In that specific sense, scientists do understand how the network works.

Richards, Kording, and their coauthors argue that this sort of understanding could carry over into neuroscience. If a deep neural network trained to complete a particular task, such as object recognition, closely resembles the brain, it stands to reason that the brain might have been optimized to do something similar. Of course, human and animal brains do not learn to do specific tasks using enormous volumes of data like deep neural networks do, but they were iteratively refined over millions of years of evolution. If a deep neural network trained to do a particular task winds up closely resembling the brain in the way that DiCarlo's networks closely resemble the visual regions of the macaque brain, then perhaps both network and brain were optimized for the same purpose. The network was trained to do that task; the brain evolved to do it.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 04:10:36, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://www.cambridge.org/core/product/CFCF20125BB1ED3E1D2D8A0A3FEED7A6

When we spoke, Kording referred to this flavor of understanding as "machine understanding," in contrast to the "human understanding" afforded by mechanistic explanation. And this "machine understanding" certainly counts as a form of understanding – in providing scientists with promising theories about the brain's evolutionary history, it empowers them to answer questions about how and why the brain came to be the way that it is. But it doesn't help neuroscientists dissect the brain's mechanisms, which has long been, as Kording puts it, the dream of neuroscience.

"We want human understanding and the world gives us machine understanding," Kording says. "And it doesn't satisfy the same urges."

The Lessons of Interpretability

Neuroscientists aren't giving up entirely on mechanisms. Lindsay is relatively sanguine that we might be able to achieve mechanistic understanding of the brains of simple animals, such as fruit flies. In those animals, brain structure and function are, for the most part, genetically preprogrammed; learning doesn't reshape their brains as dramatically as it does ours. Similarly, Richards expects that tried-and-true approaches to neuroscientific understanding will prove fruitful in genetically preprogrammed areas of the human brain – such as the brain stem, which controls automatic behaviors ranging from breathing to sleep.

Nor is Lindsay giving up on understanding more complex areas of the brain, such as the visual cortex and the frontal lobe. While she is listed as a coauthor on "A Deep

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 04:10:36, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://www.cambridge.org/core/product/CFCF20125BB1ED3E1D2D8A0A3FEED7A6

Learning Framework for Neuroscience," she says that's simply because she participated in the meeting at which its key ideas were formulated, and that she was a "dissenting voice" in those discussions. "I do think it's a little bit giving up," she says. "It's like, 'You know what, it's a mess in there, so let's not worry about trying to make sense of it.'" She thinks the field should still strive to build human-interpretable accounts of how particular brain areas work – which, she says, may require developing new mathematical concepts and tools to analyze the large-scale data and complex models that have become the bread and butter of her discipline.

That said, she's not ready to wager her life savings on the future of mechanistic understanding in neuroscience. "I'm very uncertain about all of this," she says. "I'm putting out a hopeful vision of what neuroscience could one day be, but I don't know what's going to happen and if it's possible to understand the brain."

When I spoke with Richards, who continues to advocate for the ideas expressed in "A Deep Learning Framework for Neuroscience," he shared a perspective that was surprisingly similar to Lindsay's. He doesn't discount the possibility that future tools and techniques could conceivably help neuroscientists build a mechanistic account of how the brain works, and he agrees that such an account would be desirable. But he also highlights the possibility that the brain might be too complicated an object for our limited cognitive capacities to handle, no matter how sophisticated our mathematical and experimental approaches become. "The human brain literally might not be capable of understanding [itself]," he says.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 04:10:36, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://www.cambridge.org/core/product/CFCF20125BB1ED3E1D2D8A0A3FEED7A6

As neuroscientists work to advance their techniques for understanding how the brain works, they might draw inspiration from mechanistic interpretability researchers, who grapple with a similar set of questions. Mechanistic interpretability is often compared to neuroscience, and with good reason: Researchers in the field study neural networks much as neuroscientists study brains. They track the behavior of various neurons as networks run, observe the effects of strongly activating or totally deactivating those neurons, and map out how those neurons work together in larger circuits. The goal, as in neuroscience, is to uncover mechanistic explanations for how the network functions.

Because of the similarities between brains and deep neural networks, strategies that help researchers understand how artificial neural networks work might also prove fruitful for neuroscientists. Like golfers practicing their swing on an indoor putting green, mechanistic interpretability researchers can rapidly test and refine their approaches in simulated networks before neuroscientists carry them over into the more challenging domain of the biological brain. The rigors of studying biological networks, however, are bracing: Whereas mechanistic interpretability researchers can instantly view or modify the activity of any neuron in their network, neuroscientists have to resign themselves to accessing only a tiny fraction of the brain's neurons in any given experiment. If studying artificial neural networks is like reading Shel Silverstein, studying brain networks in living animals is more like piecing together a Sapphic ode from papyrus fragments preserved in an ancient trash heap.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 04:10:36, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://www.cambridge.org/core/product/CFCF20125BB1ED3E1D2D8A0A3FEED7A6

Fortunately, it might be possible to apply mechanistic interpretability approaches to neuroscience without grappling with the messiness and obscurity of biology. Thanks to deep neural networks, we now have solid computer models of various brain regions – such as the networks that DiCarlo used to model how vision works in the macaque brain. If mechanistic interpretability research advances far enough that scientists can formulate mechanistic explanations for how those models work, the explanations would also be solid hypotheses about how the brain itself works. Using such proxies is common practice in biology: Studying the human body is difficult and often ethically dubious, so many scientists use model organisms, such as mice and monkeys, to formulate specific hypotheses about what is going on in humans. Deep neural networks could similarly function as "model organisms" to help neuroscientists indirectly study the human brain.

There's just one issue: Despite the support of AI companies with deep pockets, and despite the meaningful progress that mechanistic interpretability researchers have made, we still aren't anywhere close to understanding how deep neural networks function.

For neuroscientists, that's a disheartening sign. If a neuroscientist wants to observe the activity of a neuron in an animal, they must anesthetize the animal, open up its skull, carefully insert an electrode, secure the electrode, wait for the animal to heal, and pray that nothing goes wrong. A mechanistic interpretability researcher just needs to write a line of code. Even with this enormous advantage, mechanistic interpretability researchers still haven't been able to explain how any deep neural network works in its entirety.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 04:10:36, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://www.cambridge.org/core/product/CFCF20125BB1ED3E1D2D8A0A3FEED7A6

In a sense, the field of mechanistic interpretability makes the same argument as Jonas and Kording's computer chip paper, albeit at a much larger scale. Even with generous resources and all the data that one could ever want, mechanistic interpretability researchers still struggle to find effective approaches to understanding neural networks, whether artificial or biological. "I think having [deep neural network] models should be making a lot of neuroscientists question what they're doing and how they're doing it, and if we really have the ability to make progress," Lindsay says.

Richards agrees. "In neuroscience, it's even more hopeless," he says.

Nevertheless, several of the neuroscientists I spoke with believe that collecting more data, and designing better tools for collecting that data, should be a priority for the field. Last year, Richards and Eva Dyer published an article exhorting their peers to "embrace the brain's complexity" and prioritize accumulating large volumes of data.[20] The illusion that amassing enormous volumes of data would automatically unlock the brain's secrets is long shattered – and yet, as deep neural networks take on an ever-more central role in neuroscientific research, data has never been more valuable.

Recently, scientists have been using large-scale neuroscientific data to train foundation models that learn general patterns about how the neurons of animals of a particular species behave in specific situations.[21] Once such models exist, scientists can retrain them to predict how an individual animal's neurons will respond in a wide variety of contexts, using only a small amount of data from that individual. If such models work well enough, neuroscientists

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 04:10:36, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://www.cambridge.org/core/product/CFCF20125BB1ED3E1D2D8A0A3FEED7A6

could run experiments directly on the computer models instead of using living animals. The hope is that doing so could dramatically speed up progress in neuroscience.

Even simpler deep neural networks, such as the ones that DiCarlo used to model the macaque visual system, might be improved if researchers got access to more neural data. Though the simulated neurons in these networks often do an impressive job of capturing how an animal's neurons behave while performing a specific task, the networks are still highly unrealistic. For example, in real brains, visual information flows in two directions – inputs are carried forward from the retinas to the visual cortex and then to higher brain regions, and signals from those higher regions are also carried backward to modulate how the sensory information is processed. But in deep neural networks, information only flows forward. Collecting more data about the brain's anatomy and physiology could help scientists figure out how to close the gap between artificial and biological neural networks.

The upsides of building more faithful neural network models of the brain could be enormous. Imagine, for instance, a deep neural network that could predict how the human brain – or, even better, a specific person's brain – might respond to a wide range of medications, procedures, and experiences. We're nowhere near there yet, but such a tool, if it did exist, would revolutionize medicine in general and psychiatry in particular. As things currently stand, devising treatment plans for mental illness is often a matter of trial and error, and patients can wait months or years before seeing meaningful improvement. A deep neural

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 04:10:36, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://www.cambridge.org/core/product/CFCF20125BB1ED3E1D2D8A0A3FEED7A6

network that predicts nearly instantaneously how someone's brain would respond to various treatments would remove the guesswork entirely. The potential human impact is vast.

What Do We Lose?

It's not tough to make the case for deep neural networks in neuroscience. If you're feeling particularly ambitious, you could detail how simulating human brains might someday revolutionize psychiatry; if you want to stay grounded in the current evidence, you might mention the half-dozen people who have regained the ability to speak thanks to deep neural networks that decode their brain activity.

But as the value of deep neural networks has grown clearer, it has become harder to advocate for understanding the brain. Back in the day, neuroscientific understanding seemed to come with clear practical benefits: Chirimuuta, the Edinburgh philosopher, says that her doctoral advisor would justify his visual system research in grant applications by describing how it would help engineers build computer systems that could make sense of images. "The whole motivation for what we were doing as vision scientists was to figure out how the visual cortex works so that people [could] build machine vision systems," she says.

But thanks to deep neural networks, we now do have effective computer vision systems – and we still don't fully understand the visual cortex.

That said, it's still too early to say whether the arrival of deep neural networks will help or hinder the quest to

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 04:10:36, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://www.cambridge.org/core/product/CFCF20125BB1ED3E1D2D8A0A3FEED7A6

understand the brain. On the one hand, the practical capabilities of deep neural networks do undermine some justifications for understanding-oriented research, such as the argument deployed by Chirimuuta's supervisor. On the other hand, deep neural networks might prove a boon for neuroscience research in general. One clear example is the nascent field of NeuroAI, which focuses on both deploying AI tools to accelerate neuroscience and leveraging neuroscientific insights to improve AI.

There are good reasons to think that neuroscience could advance AI research – after all, deep neural networks were originally inspired by neuroscientific discoveries. Some researchers, including Dyer, hope that insights gleaned from studying the comparatively efficient brain could help engineers reduce AI's gargantuan energy footprint. So as huge floods of money surge into AI research, neuroscientists can reasonably expect to receive at least a trickle.

At the same time, deep neural networks have forced neuroscientists to confront the difficulty of understanding the brain. They are the best brain models that we currently have, and they're so tough to figure out that they are widely referred to as black boxes. And studying them is child's play compared to researching the real brain. When physicists train deep neural networks, they can choose to avoid the messy, often unsatisfying task of probing their inner workings, because they care more about what the model has learned about the world than the specific mechanisms that it uses to make its predictions. But when neuroscientists use deep neural network models of the brain, those mechanisms matter: They represent a hypothesis about how the brain

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 04:10:36, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://www.cambridge.org/core/product/CFCF20125BB1ED3E1D2D8A0A3FEED7A6

works. Perhaps that's why some neuroscientists are speaking much more directly about abandoning understanding – or at least their old ideals of understanding – than scientists in other fields. When the jet-black impenetrability of deep neural networks stares them directly in the face, they don't have the option to look away.

Deep neural networks can't entirely be to blame for the movement away from mechanistic understanding in neuroscience. The brain is vexingly complicated, and though deep neural networks have surfaced that complexity, they are not responsible for it. But the current state of neuroscience does provide a preview of what it might look like to relinquish or evolve our ideals of understanding in other areas of science.

And it doesn't look easy. "If you talk with scientists, they're always conflicted," Kording says. The drive to deploy new technologies to improve the lives of those with neurological and psychiatric conditions is strong – but so too is the drive to understand how the brain works, neuron by neuron and circuit by circuit.

All scientists experience the drive to understand, but it may be particularly acute in neuroscience. When we grasp something about the brain – how cells in the occipital lobe detect edges, or how dopamine-releasing neurons respond to signals of future rewards – we're not just furthering our understanding of how one of the most complex objects in the universe works: We're also furthering our understanding of ourselves.

In some cases, the self-understanding afforded by neuroscience can make a real difference to people's lives.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 04:10:36, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://www.cambridge.org/core/product/CFCF20125BB1ED3E1D2D8A0A3FEED7A6

Some medications and brain injuries make people behave in new and upsetting ways – people who take drugs to treat their Parkinson's disease can develop gambling addictions, and there are rare forms of brain damage that lead to sexual compulsions. Neuroscience can help them understand that these behaviors are not moral failings but rather the predictable effects of physical changes in their brains.

More often, though, the answers provided by neuroscience don't come with any clear moral upside; they simply satisfy our basic curiosity about ourselves. That curiosity might have a different character from our curiosity about the puzzle of protein folding or the nature of the universe, but it's the same kind of need. Scientific explanations are like good art: One could live a healthy and happy life without either, but that life would be thin in places. There's an argument to be made that scientific understanding should be considered an aesthetic value.

And aesthetic values do matter. But in the face of key ethical imperatives – restoring someone's ability to communicate, curing a psychiatric disease – they can be tough to argue for. "I don't say that [understanding is] a bad thing, just like I wouldn't say beautiful art is a bad thing," DiCarlo told me. "But I would also say the taxpayers and the public think they're paying us for something more than that."

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 04:10:36, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://www.cambridge.org/core/product/CFCF20125BB1ED3E1D2D8A0A3FEED7A6