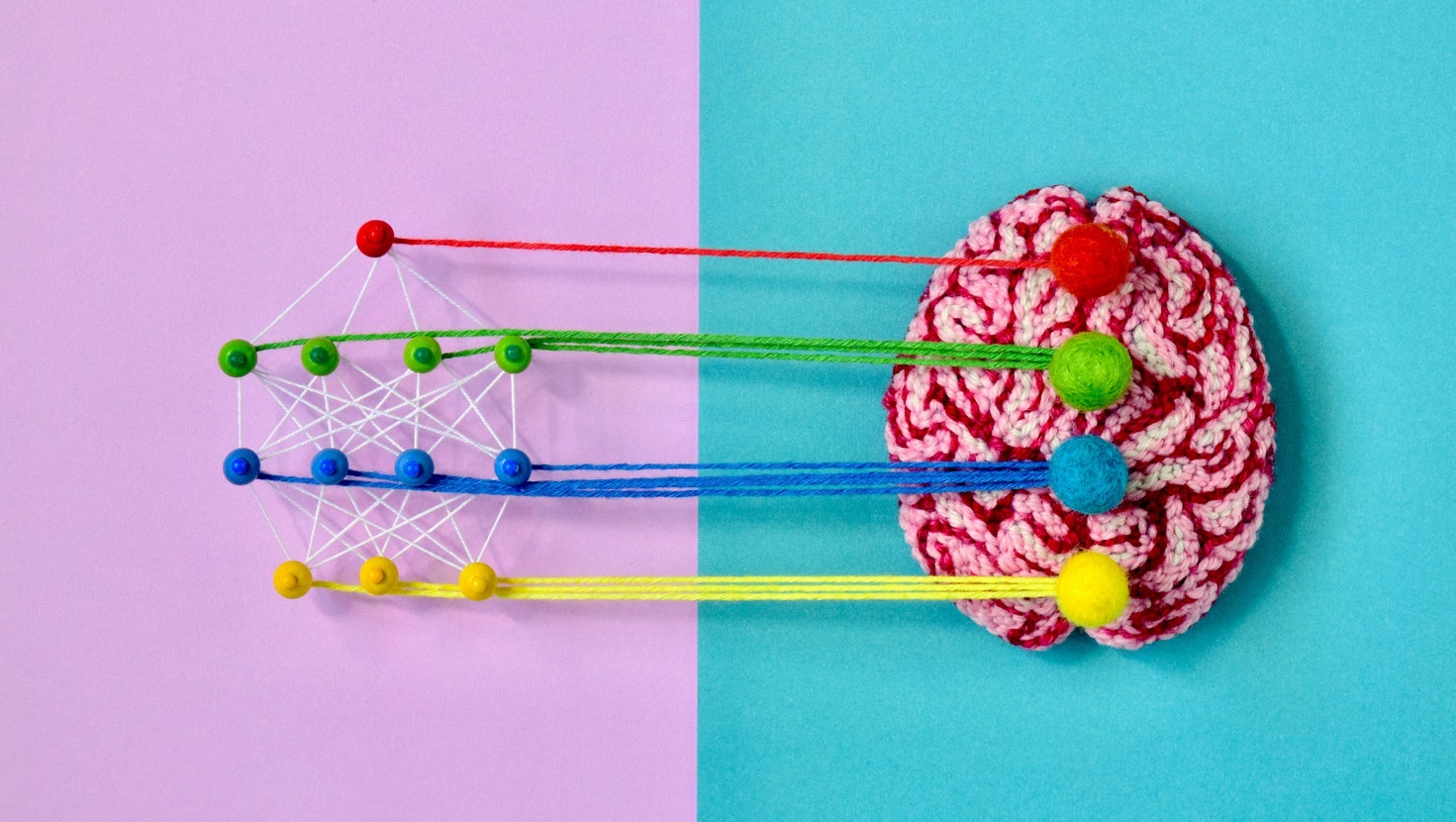
**Deep Neural Networks Are Helping Decipher How Brains Work**

**Neuroscientists are finding that deep-learning networks, often criticized as “black boxes,” can be good models for the organization of living brains.**

Computational neuroscientists are finding that deep-learning neural networks can be good explanatory models for the functional organization of living brains.PHOTOGRAPH: HINÉ MIZUSHIMA/QUANTA MAGAZINE

**IN THE WINTER** of 2011, Daniel Yamins, a postdoctoral researcher in computational neuroscience at the Massachusetts Institute of Technology, would at times toil past midnight on his machine vision project. He was painstakingly designing a system that could recognize objects in pictures, regardless of variations in size, position, and other properties—something that humans do with ease. The system was a deep neural network, a type of computational device inspired by the neurological wiring of living brains.

[Original story](https://www.quantamagazine.org/deep-neural-networks-help-to-explain-living-brains-20201028/) reprinted with permission from [*Quanta Magazine*](https://www.quantamagazine.org/), an editorially independent publication of the [Simons Foundation](https://www.simonsfoundation.org/) whose mission is to enhance public understanding of science by covering research develop­ments and trends in mathe­matics and the physical and life sciences.

“I remember very distinctly the time when we found a neural network that actually solved the task,” he said. It was 2 am, a tad too early to wake up his adviser, James DiCarlo, or other colleagues, so an excited Yamins took a walk in the cold Cambridge air. “I was really pumped,” he said.

It would have counted as a noteworthy accomplishment in artificial intelligence alone, one of many that would make neural networks the darlings of AI technology over the next few years. But that wasn’t the main goal for Yamins and his colleagues. To them and other neuroscientists, this was a pivotal moment in the development of computational models for brain functions.

DiCarlo and Yamins, who now runs his own lab at Stanford University, are part of a coterie of neuroscientists using deep neural networks to make sense of the brain’s architecture. In particular, scientists have struggled to understand the reasons behind the specializations within the brain for various tasks. They have wondered not just why different parts of the brain do different things, but also why the differences can be so specific: Why, for example, does the brain have an area for recognizing objects in general but also for faces in particular? Deep neural networks are showing that such specializations may be the most efficient way to solve problems.



The computational neuroscientist Daniel Yamins, now at Stanford University, showed that a neural network processing the features of a scene hierarchically, much as the brain does, could match the performance of humans at recognizing objects.PHOTOGRAPH: FONTEJON PHOTOGRAPHY/WU TSAI NEUROSCIENCES INSTITUTE

Similarly, researchers have demonstrated that the deep networks most proficient at classifying speech, music, and simulated scents have architectures that seem to parallel the brain’s auditory and olfactory systems. Such parallels also show up in deep nets that can look at a 2D scene and infer the underlying properties of the 3D objects within it, which helps to explain how biological perception can be both fast and incredibly rich. All these results hint that the structures of living neural systems embody certain optimal solutions to the tasks they have taken on.

These successes are all the more unexpected given that neuroscientists have long been skeptical of comparisons between brains and deep neural networks, whose workings can be inscrutable. “Honestly, nobody in my lab was doing anything with deep nets [until recently],” said the MIT neuroscientist Nancy Kanwisher. “Now, most of them are training them routinely.”

**Deep Nets and Vision**

Artificial neural networks are built with interconnecting components called perceptrons, which are simplified digital models of biological neurons. The networks have at least two layers of perceptrons, one for the input layer and one for the output. Sandwich one or more “hidden” layers between the input and the output and [you get a “deep” neural network](https://www.quantamagazine.org/how-to-win-at-deep-learning-20171009/); the greater the number of hidden layers, the deeper the network.

Deep nets can be trained to pick out patterns in data, such as patterns representing the images of cats or dogs. Training involves using an algorithm to iteratively adjust the strength of the connections between the perceptrons, so that the network learns to associate a given input (the pixels of an image) with the correct label (cat or dog). Once trained, the deep net should ideally be able to classify an input it hasn’t seen before.

In their general structure and function, deep nets aspire loosely to emulate brains, in which the adjusted strengths of connections between neurons reflect learned associations. Neuroscientists have often pointed out important limitations in that comparison: [Individual neurons may process information](https://www.quantamagazine.org/neural-dendrites-reveal-their-computational-power-20200114/) more extensively than “dumb” perceptrons do, for example, and deep nets frequently depend on a kind of communication between perceptrons called back-propagation that does not seem to occur in nervous systems. Nevertheless, for computational neuroscientists, deep nets have sometimes seemed like the best available option for modeling parts of the brain.

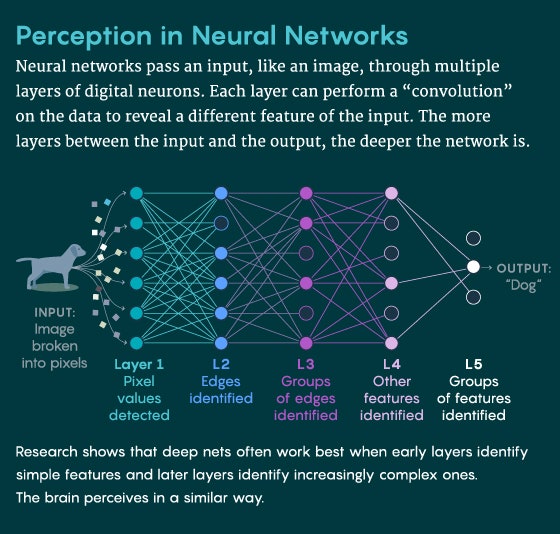


ILLUSTRATION: LUCY READING-IKKANDA/SAMUEL VELASCO/QUANTA MAGAZINE

Researchers developing computational models of the visual system have been influenced by what we know of the primate visual system, particularly the pathway responsible for recognizing people, places, and things called the ventral visual stream. (A largely separate pathway, the dorsal visual stream, processes information for seeing motion and the positions of things.) In humans, this ventral pathway begins in the eyes and proceeds to the lateral geniculate nucleus in the thalamus, a sort of relay station for sensory information. The lateral geniculate nucleus connects to an area called V1 in the primary visual cortex, downstream of which lie areas V2 and V4, which finally lead to the inferior temporal cortex. (Nonhuman primate brains have homologous structures.)

The key neuroscientific insight is that visual information processing is hierarchical and proceeds in stages: The earlier stages process low-level features in the visual field (such as edges, contours, colors and shapes), whereas complex representations, such as whole objects and faces, emerge only later in the inferior temporal cortex.

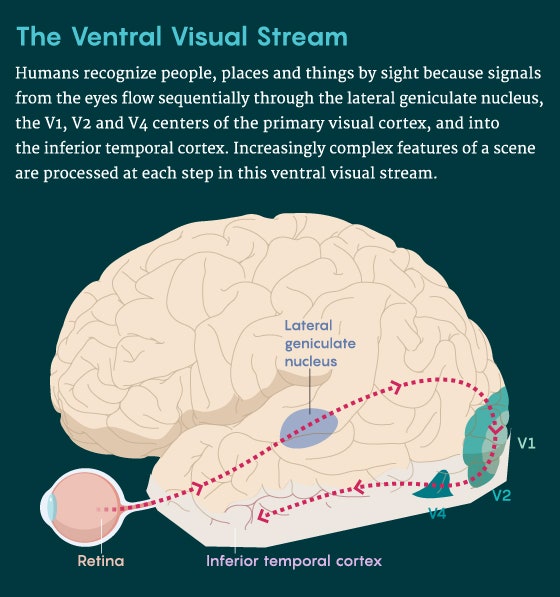


ILLUSTRATION: SAMUEL VELASCO/QUANTA MAGAZINE

Those insights guided the design of the deep net by Yamins and his colleagues. Their deep net had hidden layers, some of which performed a “convolution” that applied the same filter to every portion of an image. Each convolution captured different essential features of the image, such as edges. The more basic features were captured in the early stages of the network and the more complex features in the deeper stages, as in the primate visual system. When a convolutional neural network (CNN) like this one is trained to classify images, it starts off with randomly initialized values for its filters and learns the correct values needed for the task at hand.

The team’s four-layer CNN could recognize eight categories of objects (animals, boats, cars, chairs, faces, fruits, planes and tables) depicted in 5,760 photo-realistic 3D images. The pictured objects varied greatly in pose, position, and scale. Even so, the deep net matched the performance of humans, who are extremely good at recognizing objects despite variation.

Unbeknownst to Yamins, a revolution brewing in the world of computer vision would also independently validate the approach that he and his colleagues were taking. Soon after they finished building their CNN, another CNN called AlexNet made a name for itself at an annual image recognition contest. AlexNet, too, was based on a hierarchical processing architecture that captured basic visual features in its early stages and more complex features at higher stages; it had been trained on 1.2 million labeled images presenting a thousand categories of objects. In the 2012 contest, AlexNet routed all other tested algorithms: By the metrics of the competition, AlexNet’s error rate was only 15.3 percent, compared to 26.2 percent for its nearest competitor. With AlexNet’s victory, deep nets became legitimate contenders in the field of AI and machine learning.

Yamins and other members of DiCarlo’s team, however, were after a neuroscientific payoff. If their CNN mimicked a visual system, they wondered, could it predict neural responses to a novel image? To find out, they first established how the activity in sets of artificial neurons in their CNN corresponded to activity in almost 300 sites in the ventral visual stream of two rhesus macaques.

Then they used the CNN to predict how those brain sites would respond when the monkeys were shown images that weren’t part of the training data set. “Not only did we get good predictions … but also there’s a kind of anatomical consistency,” Yamins said: The early, intermediary, and late-stage layers of the CNN predicted the behaviors of the early, intermediary, and higher-level brain areas, respectively. Form followed function.

Kanwisher remembers being impressed by the result when it was published in 2014. “It doesn’t say that the units in the deep network individually behave like neurons biophysically,” she said. “Nonetheless, there is shocking specificity in the functional match.”

**Specializing for Sounds**

After the results from Yamins and DiCarlo appeared, the hunt was on for other, better deep-net models of the brain, particularly for regions less well studied than the primate visual system. For example, “we still don’t really have a very good understanding of the auditory cortex, particularly in humans,” said Josh McDermott, a neuroscientist at MIT. Could deep learning help generate hypotheses about how the brain processes sounds?



The neuroscientist Josh McDermott at the Massachusetts Institute of Technology uses deep learning neural networks to develop better models for auditory processing in the brain.PHOTOGRAPH: JUSTIN KNIGHT/MCGOVERN INSTITUTE

That’s McDermott’s goal. His team, which included Alexander Kell and Yamins, began designing deep nets to classify two types of sounds: speech and music. First, they hard-coded a model of the cochlea—the sound-transducing organ in the inner ear, whose workings are understood in great detail—to process audio and sort the sounds into different frequency channels as inputs to a convolutional neural network. The CNN was trained both to recognize words in audio clips of speech and to recognize the genres of musical clips mixed with background noise. The team searched for a deep-net architecture that could perform these tasks accurately without needing a lot of resources.

Three sets of architectures seemed possible. The deep net’s two tasks could share only the input layer and then split into two distinct networks. At the other extreme, the tasks could share the same network for all their processing and split only at the output stage. Or it could be one of the dozens of variants in between, where some stages of the network were shared and others were distinct.

Unsurprisingly, the networks that had dedicated pathways after the input layer outdid the networks that fully shared pathways. However, a hybrid network—one with seven common layers after the input stage and then two separate networks of five layers each—did almost as well as the fully separate network. McDermott and colleagues chose the hybrid network as the one that worked best with the least computational resources.

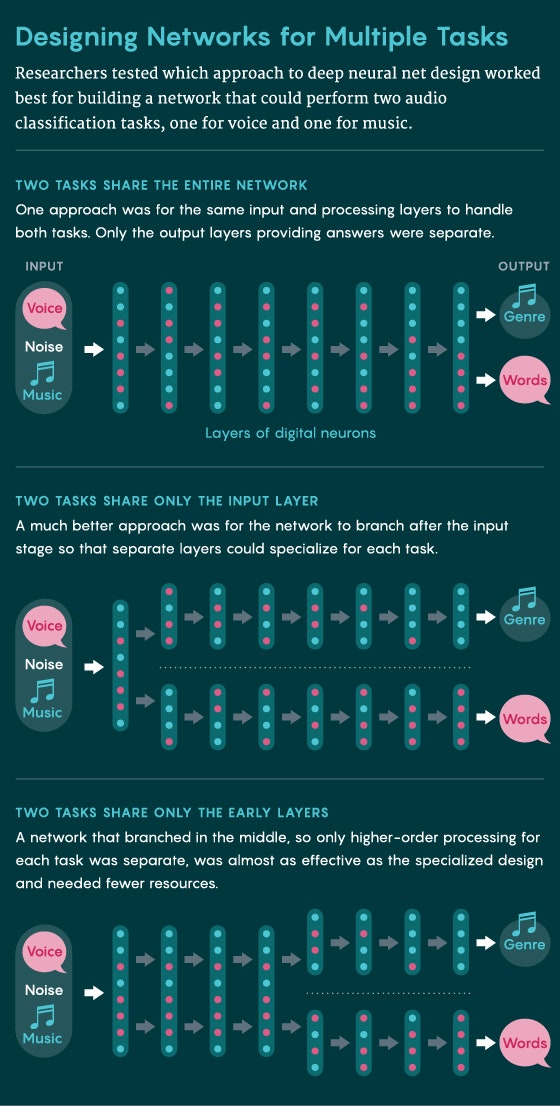


ILLUSTRATION: SAMUEL VELASCO/QUANTA MAGAZINE

When they pitted that hybrid network against humans in these tasks, it matched up well. It also matched up to earlier results from a number of researchers that suggested the non-primary auditory cortex has distinct regions for processing music and speech. And in a key test published in 2018, the model predicted the brain activity in human subjects: The model’s intermediate layers anticipated the responses of the primary auditory cortex, and deeper layers anticipated higher areas in the auditory cortex. These predictions were substantially better than those of models not based on deep learning.

“The goal of the science is to be able to predict what systems are going to do,” said McDermott. “These artificial neural networks get us closer to that goal in neuroscience.”

Kanwisher, initially skeptical of deep learning’s usefulness for her own research, was inspired by McDermott’s models. Kanwisher is best known for her work in the mid- to late 1990s showing that a region of the inferior temporal cortex called the fusiform face area (FFA) is specialized for the identification of faces. The FFA is significantly more active when subjects stare at images of faces than when they’re looking at images of objects such as houses. Why does the brain segregate the processing of faces from that of other objects?

Traditionally, answering such “why” questions has been hard for neuroscience. So Kanwisher, along with her postdoc Katharina Dobs and other colleagues, turned to deep nets for help. They used a computer-vision successor to AlexNet—a much deeper convolutional neural network called VGG—and trained two separate deep nets in specific tasks: recognizing faces and recognizing objects.



Alexander Kell, now a postdoctoral researcher at Columbia University, worked with McDermott at MIT on evaluating the effectiveness of different architectural strategies in the design of neural nets that performed multiple auditory tasks.COURTESY OF ALEX KELL

The team found that the deep net trained to recognize faces was bad at recognizing objects and vice versa, suggesting that these networks represent faces and objects differently. Next, the team trained a single network on both tasks. They found that the network had internally organized itself to segregate the processing of faces and objects in the later stages of the network. “VGG spontaneously segregates more at the later stages,” Kanwisher said. “It doesn’t have to segregate at the earlier stages.”

This agrees with the way the human visual system is organized: Branching happens only downstream of the shared earlier stages of the ventral visual pathway (the lateral geniculate nucleus and areas V1 and V2). “We found that functional specialization of face and object processing spontaneously emerged in deep nets trained on both tasks, like it does in the human brain,” said Dobs, who is now at Justus Liebig University in Giessen, Germany.

“What’s most exciting to me is that I think we have now a way to answer questions about why the brain is the way it is,” Kanwisher said.

**Layers of Scents**

More such evidence is emerging from research tackling the perception of smells. Last year, the computational neuroscientist Robert Yang and his colleagues at Columbia University designed a deep net to model the olfactory system of a fruit fly, which has been mapped in great detail by neuroscientists.

The first layer of odor processing involves olfactory sensory neurons, each of which expresses only one of about 50 types of odor receptors. All the sensory neurons of the same type, about 10 on average, reach out to a single nerve cluster in the next layer of the processing hierarchy. Because there are about 50 such nerve clusters on each side of the brain in this layer, this establishes a one-to-one mapping between types of sensory neurons and corresponding nerve clusters. The nerve clusters have multiple random connections to neurons in the next layer, called the Kenyon layer, which has about 2,500 neurons, each of which receives about seven inputs. The Kenyon layer is thought to be involved in high-level representations of the odors. A final layer of about 20 neurons provides the output that the fly uses to guide its smell-related actions (Yang cautions that no one knows whether this output qualifies as classification of odors).

To see if they could design a computational model to mimic this process, Yang and colleagues first created a data set to mimic smells, which don’t activate neurons in the same way as images. If you superimpose two images of cats, adding them pixel by pixel, the resulting image may look nothing like a cat. However, if you mix an odor from two apples, it’ll likely still smell like an apple. “That’s a critical insight that we used to design our olfaction task,” said Yang. They built their deep net with four layers: three that modeled processing layers in the fruit fly and an output layer. When Yang and colleagues trained this network to classify the simulated odors, they found that the network converged on much the same connectivity as seen in the fruit fly brain: a one-to-one mapping from layer 1 to layer 2, and then a sparse and random (7-to-1) mapping from layer 2 to layer 3.

This similarity suggests that both evolution and the deep net have reached an optimal solution. But Yang remains wary about their results. “Maybe we just got lucky here, and maybe it doesn’t generalize,” he said.

The next step in testing will be to evolve deep networks that can predict the connectivity in the olfactory system of some animal not yet studied, which can then be confirmed by neuroscientists. “That will provide a much more stringent test of our theory,” said Yang, who will move to MIT in July 2021.

**Not Just Black Boxes**

Deep nets are often derided for being unable to generalize to data that strays too far from the training data set. They’re also infamous for being black boxes. It’s impossible to explain a deep net’s decisions by examining the millions or even billions of parameters shaping it. Isn’t a deep-net model of some part of the brain merely replacing one black box with another?

Not quite, in Yang’s opinion. “It’s still easier to study than the brain,” he said.

Last year, DiCarlo’s team published results that took on both the opacity of deep nets and their alleged inability to generalize. The researchers used a version of AlexNet to model the ventral visual stream of macaques and figured out the correspondences between the artificial neuron units and neural sites in the monkeys’ V4 area. Then, using the computational model, they synthesized images that they predicted would elicit unnaturally high levels of activity in the monkey neurons. In one experiment, when these “unnatural” images were shown to monkeys, they elevated the activity of 68 percent of the neural sites beyond their usual levels; in another, the images drove up activity in one neuron while suppressing it in nearby neurons. Both results were predicted by the neural-net model.

To the researchers, these results suggest that the deep nets do generalize to brains and are not entirely unfathomable. “However, we acknowledge that … many other notions of ‘understanding’ remain to be explored to see whether and how these models add value,” they wrote.

The convergences in structure and performance between deep nets and brains do not necessarily mean that they work the same way; there are ways in which they demonstrably do not. But it may be that there are enough similarities for both types of systems to follow the same broad governing principles.

**Limitations of the Models**

McDermott sees potential therapeutic value in these deep-net studies. Today, when people lose hearing, it’s usually due to changes in the ear. The brain’s auditory system has to cope with the impaired input. “So if we had good models of what the rest of the auditory system was doing, we would have a better idea of what to do to actually help people hear better,” McDermott said.

Still, McDermott is cautious about what the deep nets can deliver. “We have been pushing pretty hard to try to understand the limitations of neural networks as models,” he said.



Jenelle Feather, a graduate student in McDermott’s laboratory, has used carefully designed pairs of audio inputs called metamers to compare the performance of neural networks with that of human hearing.PHOTOGRAPH: CAITLIN CUNNINGHAM/MCGOVERN INSTITUTE

In one striking demonstration of those limitations, the graduate student Jenelle Feather and others in McDermott’s lab focused on metamers, which are physically distinct input signals that produce the same representation in a system. Two audio metamers, for example, have different wave forms but sound the same to a human. Using a deep-net model of the auditory system, the team designed metamers of natural audio signals; these metamers activated different stages of the neural network in the same way the audio clips did. If the neural network accurately modeled the human auditory system, then the metamers should sound the same, too.

But that’s not what happened. Humans recognized the metamers that produced the same activation as the corresponding audio clips in the early stages of the neural network. However, this did not hold for metamers with matching activations in the deeper stages of the network: those metamers sounded like noise to humans. “So even though under certain circumstances these kinds of models do a very good job of replicating human behavior, there’s something that’s very wrong about them,” McDermott said.

At Stanford, Yamins is exploring ways in which these models are not yet representative of the brain. For instance, many of these models need loads of labeled data for training, while our brains can learn effortlessly from as little as one example. Efforts are underway to develop unsupervised deep nets that can learn as efficiently. Deep nets also learn using an algorithm called back propagation, which most neuroscientists think cannot work in real neural tissue because it lacks the appropriate connections. “There’s been some big progress made in terms of somewhat more biologically plausible learning rules that actually do work,” Yamins said.

Josh Tenenbaum, a cognitive neuroscientist at MIT, said that while all these deep-net models are “real steps of progress,” they are mainly doing classification or categorization tasks. Our brains, however, do much more than categorize what’s out there. Our vision system can make sense of the geometry of surfaces and the 3D structure of a scene, and it can reason about underlying causal factors—for example, it can infer in real time that a tree has disappeared only because a car has passed in front of it.

To understand this ability of the brain, Ilker Yildirim, formerly at MIT and now at Yale University, worked with Tenenbaum and colleagues to build something called an efficient inverse graphics model. It begins with parameters that describe a face to be rendered on a background, such as its shape, its texture, the direction of lighting, the head pose and so on. A computer graphics program called a generative model creates a 3D scene from the parameters; then, after various stages of processing, it produces a 2D image of that scene as viewed from a certain position. Using the 3D and 2D data from the generative model, the researchers trained a modified version of AlexNet to predict the likely parameters of a 3D scene from an unfamiliar 2D image. “The system learns to go backwards from the effect to the cause, from the 2D image to the 3D scene that produced it,” said Tenenbaum.

The team tested their model by verifying its predictions about activity in the inferior temporal cortex of rhesus macaques. They presented macaques with 175 images, showing 25 individuals in seven poses, and recorded the neural signatures from “face patches,” visual processing areas that specialize in face recognition. They also showed the images to their deep learning network. In the network, the activation of the artificial neurons in the first layer represents the 2D image and the activation in the last layer represents the 3D parameters. “Along the way, it goes through a bunch of transformations, which seem to basically get you from 2D to 3D,” Tenenbaum said. They found that the last three layers of the network corresponded remarkably well to the last three layers of the macaques’ face processing network.

This suggests that brains use combinations of generative and recognition models not just to recognize and characterize objects but to infer the causal structures inherent in scenes, all in an instant. Tenenbaum acknowledges that their model doesn’t prove that the brain works this way. “But it does open the door to asking those questions in a more fine-grained mechanistic way,” he said. “It should be … motivating us to walk through it.”

*Editor’s note: Daniel Yamins and James DiCarlo receive research funding from the*[*Simons Collaboration on the Global Brain*](https://www.simonsfoundation.org/collaborations/global-brain/)*, which is part of the Simons Foundation, the organization that also funds this editorially independent magazine. Simons Foundation funding decisions have no bearing on Quanta’s coverage. Please see*[*this page*](https://www.quantamagazine.org/about/)*for more details.*

[*Original story*](https://www.quantamagazine.org/deep-neural-networks-help-to-explain-living-brains-20201028/)*reprinted with permission from* [Quanta Magazine](https://www.quantamagazine.org/), *an editorially independent publication of the*[*Simons Foundation*](https://www.simonsfoundation.org/)*whose mission is to enhance public understanding of science by covering research developments and trends in mathematics and the physical and life sciences.*