PAUL VOOSEN

SHARE:

- Twitter
- Linked In
- Facebook
- Reddit
- Wechat
- Email

Jason Yosinski sits in a small glass box at Uber's San Francisco, California, headquarters, pondering the mind of an artificial intelligence. An Uber research scientist, Yosinski is performing a kind of brain surgery on the AI running on his laptop. Like many of the AIs that will soon be powering so much of modern life, including self-driving Uber cars, Yosinski's program is a deep neural network, with an architecture loosely inspired by the brain. And like the brain, the program is hard to understand from the outside: It's a black box.

This particular AI has been trained, using a vast sum of labeled images, to recognize objects as random as zebras, fire trucks, and seat belts. Could it recognize Yosinski and the reporter hovering in front of the webcam? Yosinski zooms in on one of the AI's individual computational nodes—the neurons, so to speak—to see what is prompting its response. Two ghostly white ovals pop up and float on the screen. This neuron, it seems, has learned to detect the outlines of faces. "This responds to your face and my face," he says. "It responds to different size faces, different color faces."

No one trained this network to identify faces. Humans weren't labeled in its training images. Yet learn faces it did, perhaps as a way to recognize the things that tend to accompany them, such as ties and cowboy hats. The network is too complex for humans to comprehend its exact decisions. Yosinski's probe had illuminated one small part of it, but overall, it remained opaque. "We build amazing models," he says. "But we don't quite understand them. And every year, this gap is going to get a bit larger."

Each month, it seems, deep neural networks, or deep learning, as the field is also called, spread to another scientific discipline. They can predict the best way to synthesize organic molecules. They can detect genes related to autism risk. They are even changing how science itself is conducted. The AIs often succeed in what they do. But they have left scientists, whose very enterprise is founded on explanation, with a nagging question: Why, model, why?

That interpretability problem, as it's known, is galvanizing a new generation of researchers in both industry and academia. Just as the microscope revealed the cell, these researchers are crafting tools that will allow insight into the how neural networks make decisions. Some tools probe the AI without penetrating it; some are alternative algorithms that can compete with neural nets, but with more transparency; and some use still more deep learning to get inside the black box. Taken together, they add up to a new discipline. Yosinski calls it "AI neuroscience."

Opening up the black box

Loosely modeled after the brain, deep neural networks are spurring innovation across science. But the mechanics of the models are mysterious: They are black boxes. Scientists are now developing tools to get inside the mind of the machine.

GRAPHIC: G. GRULLÓN/*SCIENCE*

Marco Ribeiro, a graduate student at the University of Washington in Seattle, strives to understand the black box by using a class of AI neuroscience tools called counter-factual probes. The idea is to vary the inputs to the AI—be they text, images, or anything else—in clever ways to see which changes affect the output, and how. Take a neural network that, for example, ingests the words of movie reviews and flags those that are positive. Ribeiro's program, called Local Interpretable Model-Agnostic Explanations (LIME), would take a review flagged as positive and create subtle variations by deleting or replacing words. Those variants would then be run through the black box to see whether it still considered them to be positive. On the basis of thousands of tests, LIME can identify the words—or parts of an image or molecular structure, or any other kind of data—most important in the AI's original judgment. The tests might reveal that the word "horrible" was vital to a panning or that "Daniel Day Lewis" led to a positive review. But although LIME can diagnose those singular examples, that result says little about the network's overall insight.

New counterfactual methods like LIME seem to emerge each month. But Mukund Sundararajan, another computer scientist at Google, devised a probe that doesn't require testing the network a thousand times over: a boon if you're trying to

understand many decisions, not just a few. Instead of varying the input randomly, Sundararajan and his team introduce a blank reference—a black image or a zeroed-out array in place of text—and transition it step-by-step toward the example being tested. Running each step through the network, they watch the jumps it makes in certainty, and from that trajectory they infer features important to a prediction.

Sundararajan compares the process to picking out the key features that identify the glass-walled space he is sitting in—outfitted with the standard medley of mugs, tables, chairs, and computers—as a Google conference room. "I can give a zillion reasons." But say you slowly dim the lights. "When the lights become very dim, only the biggest reasons stand out." Those transitions from a blank reference allow Sundararajan to capture more of the network's decisions than Ribeiro's variations do. But deeper, unanswered questions are always there, Sundararajan says—a state of mind familiar to him as a parent. "I have a 4-year-old who continually reminds me of the infinite regress of 'Why?'"

The urgency comes not just from science. According to a directive from the European Union, companies deploying algorithms that substantially influence the public must by next year create "explanations" for their models' internal logic. The Defense Advanced Research Projects Agency, the U.S. military's blue-sky research arm, is pouring $70 million into a new program, called Explainable AI, for interpreting the deep learning that powers drones and intelligence-mining operations. The drive to open the black box of AI is also coming from Silicon Valley itself, says Maya Gupta, a machine-learning researcher at Google in Mountain View, California. When she joined Google in 2012 and asked AI engineers about their problems, accuracy wasn't the only thing on their minds, she says. "I'm not sure what it's doing," they told her. "I'm not sure I can trust it."

Rich Caruana, a computer scientist at Microsoft Research in Redmond, Washington, knows that lack of trust firsthand. As a graduate student in the 1990s at Carnegie Mellon University in Pittsburgh, Pennsylvania, he joined a team trying to see whether machine learning could guide the treatment of pneumonia patients. In general, sending the hale and hearty home is best, so they can avoid picking up other infections in the hospital. But some patients, especially those with complicating factors such as asthma, should be admitted immediately. Caruana applied a neural network to a data set of symptoms and outcomes provided by 78 hospitals. It seemed to work well. But disturbingly, he saw that a simpler, transparent model trained on the same records suggested sending asthmatic patients home, indicating some flaw in the data. And he had no easy way of knowing whether his neural net had picked up the same bad lesson. "Fear of a neural net is completely justified," he says. "What really terrifies me is what else did the neural net learn that's equally wrong?"

Today's neural nets are far more powerful than those Caruana used as a graduate student, but their essence is the same. At one end sits a messy soup of data—say, millions of pictures of dogs. Those data are sucked into a network with a dozen or more computational layers, in which neuron-like connections "fire" in response to features of the input data. Each layer reacts to progressively more abstract features, allowing the final layer to distinguish, say, terrier from dachshund.

At first the system will botch the job. But each result is compared with labeled pictures of dogs. In a process called backpropagation, the outcome is sent backward through the network, enabling it to reweight the triggers for each neuron. The process repeats millions of times until the network learns—somehow—to make fine distinctions among breeds. "Using modern horsepower and chutzpah, you can get these things to really sing," Caruana says. Yet that mysterious and flexible power is precisely what makes them black boxes.

Gupta has a different tactic for coping with black boxes: She avoids them. Several years ago Gupta, who moonlights as a designer of intricate physical puzzles, began a project called GlassBox. Her goal is to tame neural networks by engineering predictability into them. Her guiding principle is monotonicity—a relationship between variables in which, all else being equal, increasing one variable directly increases another, as with the square footage of a house and its price.

Gupta embeds those monotonic relationships in sprawling databases called interpolated lookup tables. In essence, they're like the tables in the back of a high school trigonometry textbook where you'd look up the sine of 0.5. But rather than dozens of entries across one dimension, her tables have millions across multiple dimensions. She wires those tables into neural networks, effectively adding an extra, predictable layer of computation—baked-in knowledge that she says will ultimately make the network more controllable.

Caruana, meanwhile, has kept his pneumonia lesson in mind. To develop a model that would match deep learning in accuracy but avoid its opacity, he turned to a community that hasn't always gotten along with machine learning and its loosey-goosey ways: statisticians.

In the 1980s, statisticians pioneered a technique called a generalized additive model (GAM). It built on linear regression, a way to find a linear trend in a set of data. But GAMs can also handle trickier relationships by finding multiple operations that together can massage data to fit on a regression line: squaring a set of numbers while taking the logarithm for another group of variables, for example. Caruana has supercharged the process, using machine learning to discover those operations—which

can then be used as a powerful pattern-detecting model. "To our great surprise, on many problems, this is very accurate," he says. And crucially, each operation's influence on the underlying data is transparent.

Caruana's GAMs are not as good as AIs at handling certain types of messy data, such as images or sounds, on which some neural nets thrive. But for any data that would fit in the rows and columns of a spreadsheet, such as hospital records, the model can work well. For example, Caruana returned to his original pneumonia records. Reanalyzing them with one of his GAMs, he could see why the AI would have learned the wrong lesson from the admission data. Hospitals routinely put asthmatics with pneumonia in intensive care, improving their outcomes. Seeing only their rapid improvement, the AI would have recommended the patients be sent home. (It would have made the same optimistic error for pneumonia patients who also had chest pain and heart disease.)

Caruana has started touting the GAM approach to California hospitals, including Children's Hospital Los Angeles, where about a dozen doctors reviewed his model's results. They spent much of that meeting discussing what it told them about pneumonia admissions, immediately understanding its decisions. "You don't know much about health care," one doctor said, "but your model really does."

Sometimes, you have to embrace the darkness. That's the theory of researchers pursuing a third route toward interpretability. Instead of probing neural nets, or avoiding them, they say, the way to explain deep learning is simply to do more deep learning.

If we can't ask ... why they do something and get a reasonable response back, people will just put it back on the shelf.

- MARK RIEDL, GEORGIA INSTITUTE OF TECHNOLOGY

Like many AI coders, Mark Riedl, director of the Entertainment Intelligence Lab at the Georgia Institute of Technology in Atlanta, turns to 1980s video games to test his creations. One of his favorites is Frogger, in which the player navigates the eponymous amphibian through lanes of car traffic to an awaiting pond. Training a neural network to play expert Frogger is easy enough, but explaining what the AI is doing is even harder than usual.

Instead of probing that network, Riedl asked human subjects to play the game and to describe their tactics aloud in real time. Riedl recorded those comments alongside the

frog's context in the game's code: "Oh, there's a car coming for me; I need to jump forward." Armed with those two languages—the players' and the code—Riedl trained a second neural net to translate between the two, from code to English. He then wired that translation network into his original game-playing network, producing an overall AI that would say, as it waited in a lane, "I'm waiting for a hole to open up before I move." The AI could even sound frustrated when pinned on the side of the screen, cursing and complaining, "Jeez, this is hard."

Riedl calls his approach "rationalization," which he designed to help everyday users understand the robots that will soon be helping around the house and driving our cars. "If we can't ask a question about why they do something and get a reasonable response back, people will just put it back on the shelf," Riedl says. But those explanations, however soothing, prompt another question, he adds: "How wrong can the rationalizations be before people lose trust?"

Back at Uber, Yosinski has been kicked out of his glass box. Uber's meeting rooms, named after cities, are in high demand, and there is no surge pricing to thin the crowd. He's out of Doha and off to find Montreal, Canada, unconscious pattern recognition processes guiding him through the office maze—until he gets lost. His image classifier also remains a maze, and, like Riedl, he has enlisted a second AI to help him understand the first one.

Researchers have created neural networks that, in addition to filling gaps left in photos, can identify flaws in an artificial intelligence.

PHOTOS: ANH NGUYEN

First, Yosinski rejiggered the classifier to produce images instead of labeling them. Then, he and his colleagues fed it colored static and sent a signal back through it to request, for example, "more volcano." Eventually, they assumed, the network would shape that noise into its idea of a volcano. And to an extent, it did: That volcano, to human eyes, just happened to look like a gray, featureless mass. The AI and people saw differently.

Next, the team unleashed a generative adversarial network (GAN) on its images. Such AIs contain two neural networks. From a training set of images, the "generator" learns rules about imagemaking and can create synthetic images. A second "adversary" network tries to detect whether the resulting pictures are real or fake, prompting the generator to try again. That back-and-forth eventually results in crude images that contain features that humans can recognize.

Yosinski and Anh Nguyen, his former intern, connected the GAN to layers inside their original classifier network. This time, when told to create "more volcano," the GAN took the gray mush that the classifier learned and, with its own knowledge of picture structure, decoded it into a vast array of synthetic, realistic-looking volcanoes. Some dormant. Some erupting. Some at night. Some by day. And some, perhaps, with flaws—which would be clues to the classifier's knowledge gaps.

Their GAN can now be lashed to any network that uses images. Yosinski has already used it to identify problems in a network trained to write captions for random images. He reversed the network so that it can create synthetic images for any random caption input. After connecting it to the GAN, he found a startling omission. Prompted to imagine "a bird sitting on a branch," the network—using instructions translated by the GAN—generated a bucolic facsimile of a tree and branch, but with no bird. Why? After feeding altered images into the original caption model, he realized that the caption writers who trained it never described trees and a branch without involving a bird. The AI had learned the wrong lessons about what makes a bird. "This hints at what will be an important direction in AI neuroscience," Yosinski says. It was a start, a bit of a blank map shaded in.

The day was winding down, but Yosinski's work seemed to be just beginning. Another knock on the door. Yosinski and his AI were kicked out of another glass box conference room, back into Uber's maze of cities, computers, and humans. He didn't get lost this time. He wove his way past the food bar, around the plush couches, and through the exit to the elevators. It was an easy pattern. He'd learn them all soon.