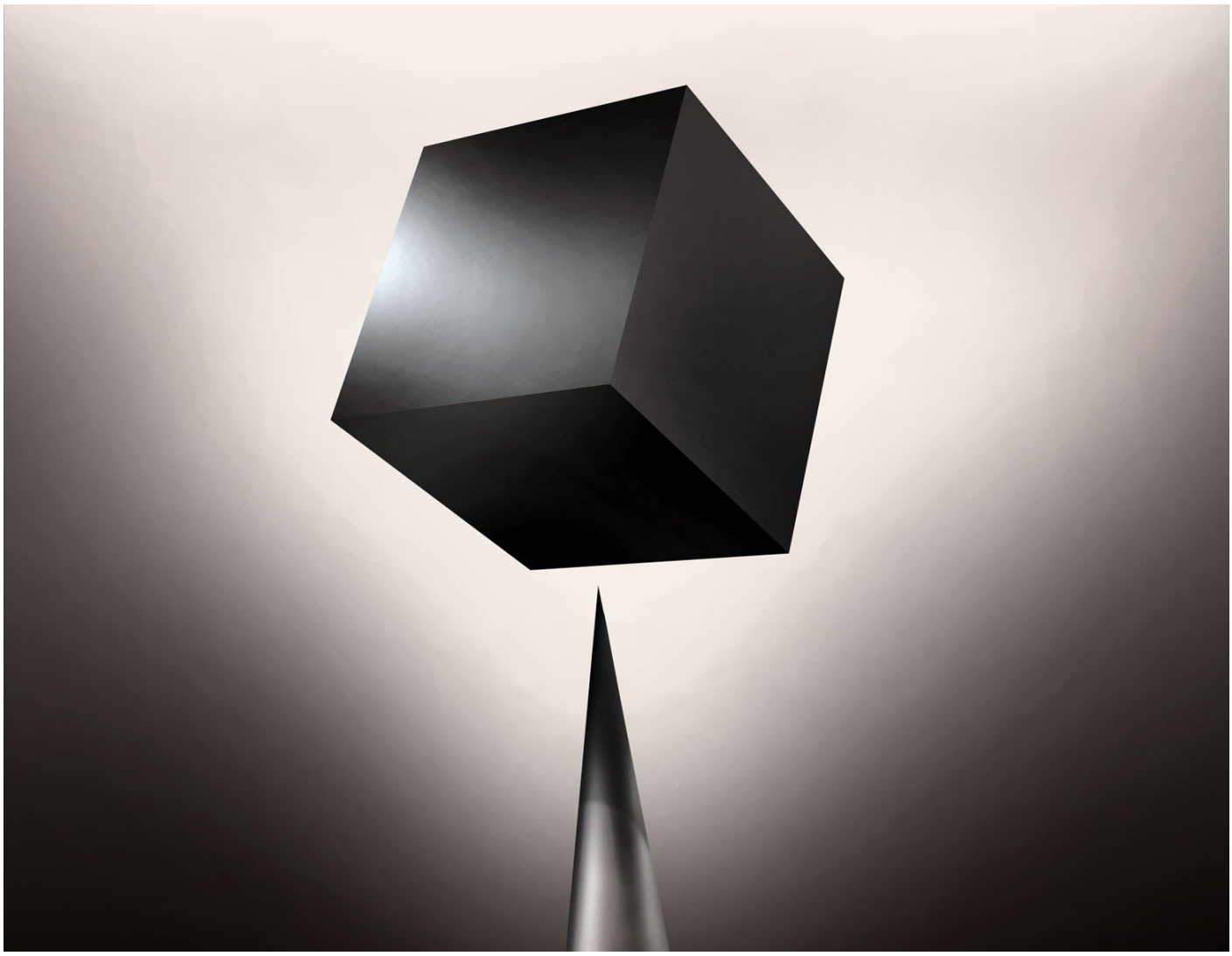


KEITH RANKIN

Intelligent Machines

The Dark Secret at the Heart of AI

No one really knows how the most advanced algorithms do what they do. That could be a problem.

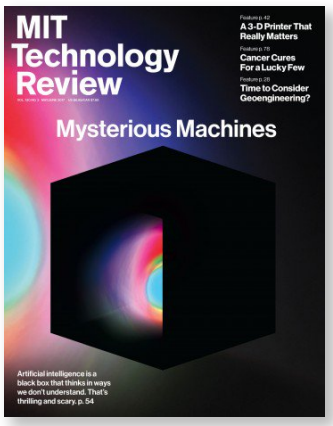
by Will Knight April 11, 2017

Last year, a strange self-driving car was released onto the quiet roads of Monmouth County, New Jersey. The experimental vehicle, developed by researchers at the chip maker Nvidia, didn't look different from other autonomous cars, but it was unlike anything demonstrated by Google, Tesla, or General Motors, and it showed the rising power of artificial intelligence. The car didn't follow a single instruction provided by an

engineer or programmer. Instead, it relied entirely on an algorithm that had taught itself to drive by watching a human do it.

Getting a car to drive this way was an impressive feat. But it's also a bit unsettling, since it isn't completely clear how the car makes its decisions. Information from the vehicle's sensors goes straight into a huge network of artificial neurons that process the data and then deliver the commands required to operate the steering wheel, the brakes, and other systems. The result seems to match the responses you'd expect from a human driver. But what if one day it did something unexpected—crashed into a tree, or sat at a green light? As things stand now, it might be difficult to find out why. The system is so complicated that even the engineers who designed it may struggle to isolate the reason for any single action. And you can't ask it: there is no obvious way to design such a system so that it could always explain why it did what it did.

The mysterious mind of this vehicle points to a looming issue with artificial intelligence. The car's underlying AI technology, known as deep learning, has proved very powerful at solving problems in recent years, and it has been widely deployed for tasks like image captioning, voice recognition, and language translation. There is now hope that the same techniques will be able to diagnose deadly diseases, make million-dollar trading decisions, and do countless other things to transform whole industries.



This story is part of our May/June 2017 Issue

See the rest of the issue

Subscribe

But this won't happen—or shouldn't happen—unless we find ways of making techniques like deep learning more understandable to their creators and accountable to their users. Otherwise it will be hard to predict when

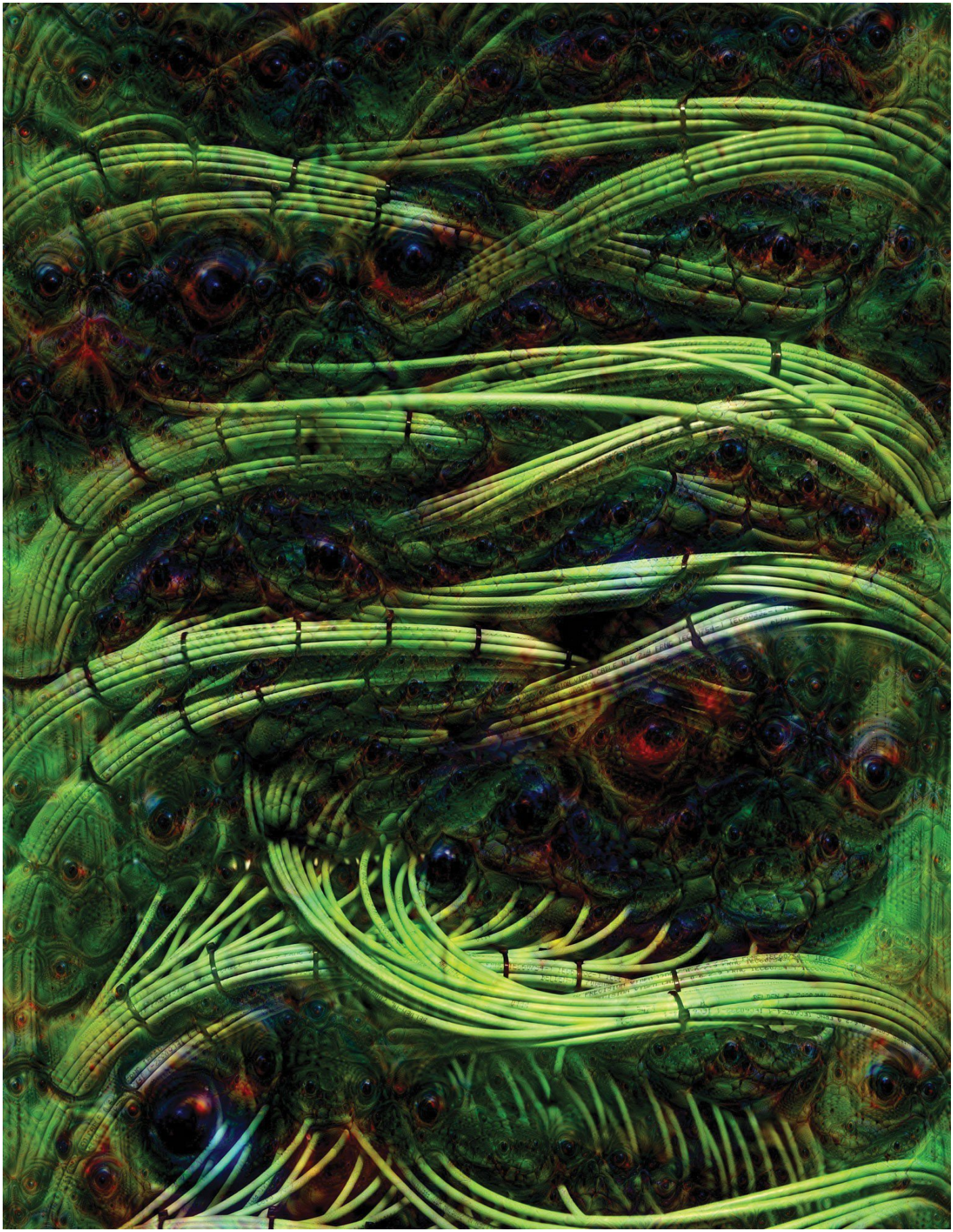
failures might occur—and it's inevitable they will. That's one reason Nvidia's car is still experimental.

Already, mathematical models are being used to help determine who makes parole, who's approved for a loan, and who gets hired for a job. If you could get access to these mathematical models, it would be possible to understand their reasoning. But banks, the military, employers, and others are now turning their attention to more complex machine-learning approaches that could make automated decision-making altogether inscrutable. Deep learning, the most common of these approaches, represents a fundamentally different way to program computers. "It is a problem that is already relevant, and it's going to be much more relevant in the future," says Tommi Jaakkola, a professor at MIT who works on applications of machine learning. "Whether it's an investment decision, a medical decision, or maybe a military decision, you don't want to just rely on a 'black box' method."

There's already an argument that being able to interrogate an AI system about how it reached its conclusions is a fundamental legal right. Starting in the summer of 2018, the European Union may require that companies be able to give users an explanation for decisions that automated systems reach. This might be impossible, even for systems that seem relatively simple on the surface, such as the apps and websites that use deep learning to serve ads or recommend songs. The computers that run those services have programmed themselves, and they have done it in ways we cannot understand. Even the engineers who build these apps cannot fully explain their behavior.

This raises mind-boggling questions. As the technology advances, we might soon cross some threshold beyond which using AI requires a leap of faith. Sure, we humans can't always truly explain our thought processes either—but we find ways to intuitively trust and gauge people. Will that also be possible with machines that think and make decisions differently from the way a human would? We've never before built machines that operate in ways their creators don't understand. How well can we expect to communicate—and get along with—intelligent machines that could be unpredictable and inscrutable? These questions took me on a journey to the

bleeding edge of research on AI algorithms, from Google to Apple and many places in between, including a meeting with one of the great philosophers of our time.



The artist Adam Ferriss created this image, and the one below, using Google Deep Dream, a program that adjusts an image to stimulate the pattern recognition capabilities of a deep neural network. The pictures were produced using a mid-level layer of the neural network.
ADAM FERRISS

In 2015, a research group at Mount Sinai Hospital in New York was inspired to apply deep learning to the hospital's vast database of patient records. This data set features hundreds of variables on patients, drawn from their test results, doctor visits, and so on. The resulting program, which the researchers named Deep Patient, was trained using data from about 700,000 individuals, and when tested on new records, it proved incredibly good at predicting disease. Without any expert instruction, Deep Patient had discovered patterns hidden in the hospital data that seemed to indicate when people were on the way to a wide range of ailments, including cancer of the liver. There are a lot of methods that are "pretty good" at predicting disease from a patient's records, says Joel Dudley, who leads the Mount Sinai team. But, he adds, "this was just way better."

"We can build these models, but we don't know how they work."

At the same time, Deep Patient is a bit puzzling. It appears to anticipate the onset of psychiatric disorders like schizophrenia surprisingly well. But since schizophrenia is notoriously difficult for physicians to predict, Dudley wondered how this was possible. He still doesn't know. The new tool offers no clue as to how it does this. If something like Deep Patient is actually going to help doctors, it will ideally give them the rationale for its prediction, to reassure them that it is accurate and to justify, say, a change in the drugs someone is being prescribed. "We can build these models," Dudley says ruefully, "but we don't know how they work."

Artificial intelligence hasn't always been this way. From the outset, there were two schools of thought regarding how understandable, or explainable, AI ought to be. Many thought it made the most sense to build machines that reasoned according to rules and logic, making their inner workings

transparent to anyone who cared to examine some code. Others felt that intelligence would more easily emerge if machines took inspiration from biology, and learned by observing and experiencing. This meant turning computer programming on its head. Instead of a programmer writing the commands to solve a problem, the program generates its own algorithm based on example data and a desired output. The machine-learning techniques that would later evolve into today's most powerful AI systems followed the latter path: the machine essentially programs itself.

At first this approach was of limited practical use, and in the 1960s and '70s it remained largely confined to the fringes of the field. Then the computerization of many industries and the emergence of large data sets renewed interest. That inspired the development of more powerful machine-learning techniques, especially new versions of one known as the artificial neural network. By the 1990s, neural networks could automatically digitize handwritten characters.

But it was not until the start of this decade, after several clever tweaks and refinements, that very large—or "deep"—neural networks demonstrated dramatic improvements in automated perception. Deep learning is responsible for today's explosion of AI. It has given computers extraordinary powers, like the ability to recognize spoken words almost as well as a person could, a skill too complex to code into the machine by hand. Deep learning has transformed computer vision and dramatically improved machine translation. It is now being used to guide all sorts of key decisions in medicine, finance, manufacturing—and beyond.

ADAM FERRISS

The workings of any machine-learning technology are inherently more opaque, even to computer scientists, than a hand-coded system. This is not

to say that all future AI techniques will be equally unknowable. But by its nature, deep learning is a particularly dark black box.

You can't just look inside a deep neural network to see how it works. A network's reasoning is embedded in the behavior of thousands of simulated neurons, arranged into dozens or even hundreds of intricately interconnected layers. The neurons in the first layer each receive an input, like the intensity of a pixel in an image, and then perform a calculation before outputting a new signal. These outputs are fed, in a complex web, to the neurons in the next layer, and so on, until an overall output is produced. Plus, there is a process known as back-propagation that tweaks the calculations of individual neurons in a way that lets the network learn to produce a desired output.

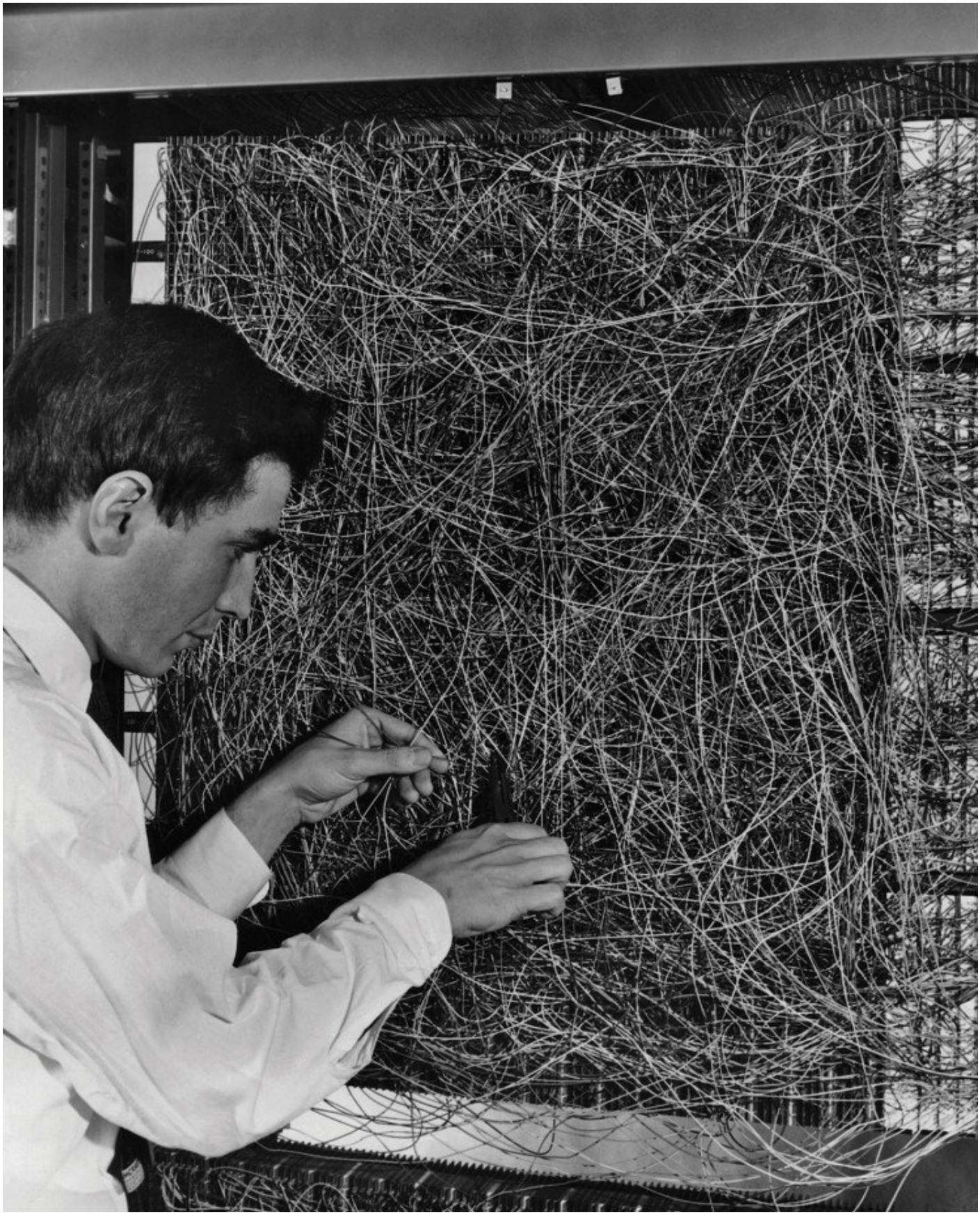
The many layers in a deep network enable it to recognize things at different levels of abstraction. In a system designed to recognize dogs, for instance, the lower layers recognize simple things like outlines or color; higher layers recognize more complex stuff like fur or eyes; and the topmost layer identifies it all as a dog. The same approach can be applied, roughly speaking, to other inputs that lead a machine to teach itself: the sounds that make up words in speech, the letters and words that create sentences in text, or the steering-wheel movements required for driving.

"It might be part of the nature of intelligence that only part of it is exposed to rational explanation. Some of it is just instinctual."

Ingenious strategies have been used to try to capture and thus explain in more detail what's happening in such systems. In 2015, researchers at Google modified a deep-learning-based image recognition algorithm so that instead of spotting objects in photos, it would generate or modify them. By effectively running the algorithm in reverse, they could discover the features

the program uses to recognize, say, a bird or building. The resulting images, produced by a project known as Deep Dream, showed grotesque, alien-like animals emerging from clouds and plants, and hallucinatory pagodas blooming across forests and mountain ranges. The images proved that deep learning need not be entirely inscrutable; they revealed that the algorithms home in on familiar visual features like a bird's beak or feathers. But the images also hinted at how different deep learning is from human perception, in that it might make something out of an artifact that we would know to ignore. Google researchers noted that when its algorithm generated images of a dumbbell, it also generated a human arm holding it. The machine had concluded that an arm was part of the thing.

Further progress has been made using ideas borrowed from neuroscience and cognitive science. A team led by Jeff Clune, an assistant professor at the University of Wyoming, has employed the AI equivalent of optical illusions to test deep neural networks. In 2015, Clune's group showed how certain images could fool such a network into perceiving things that aren't there, because the images exploit the low-level patterns the system searches for. One of Clune's collaborators, Jason Yosinski, also built a tool that acts like a probe stuck into a brain. His tool targets any neuron in the middle of the network and searches for the image that activates it the most. The images that turn up are abstract (imagine an impressionistic take on a flamingo or a school bus), highlighting the mysterious nature of the machine's perceptual abilities.



This early artificial neural network, at the Cornell Aeronautical Laboratory in Buffalo, New York, circa 1960, processed inputs from light sensors.

Ferriss was inspired to run Cornell's artificial neural network through Deep Dream, producing the images above and below.
ADAM FERRISS

We need more than a glimpse of AI's thinking, however, and there is no easy solution. It is the interplay of calculations inside a deep neural network that is crucial to higher-level pattern recognition and complex decision-making, but those calculations are a quagmire of mathematical functions and variables. "If you had a very small neural network, you might be able to understand it," Jaakkola says. "But once it becomes very large, and it has

thousands of units per layer and maybe hundreds of layers, then it becomes quite un-understandable."

In the office next to Jaakkola is Regina Barzilay, an MIT professor who is determined to apply machine learning to medicine. She was diagnosed with breast cancer a couple of years ago, at age 43. The diagnosis was shocking in itself, but Barzilay was also dismayed that cutting-edge statistical and machine-learning methods were not being used to help with oncological research or to guide patient treatment. She says AI has huge potential to revolutionize medicine, but realizing that potential will mean going beyond just medical records. She envisions using more of the raw data that she says is currently underutilized: "imaging data, pathology data, all this information."

How well can we get along with machines that are unpredictable and inscrutable?

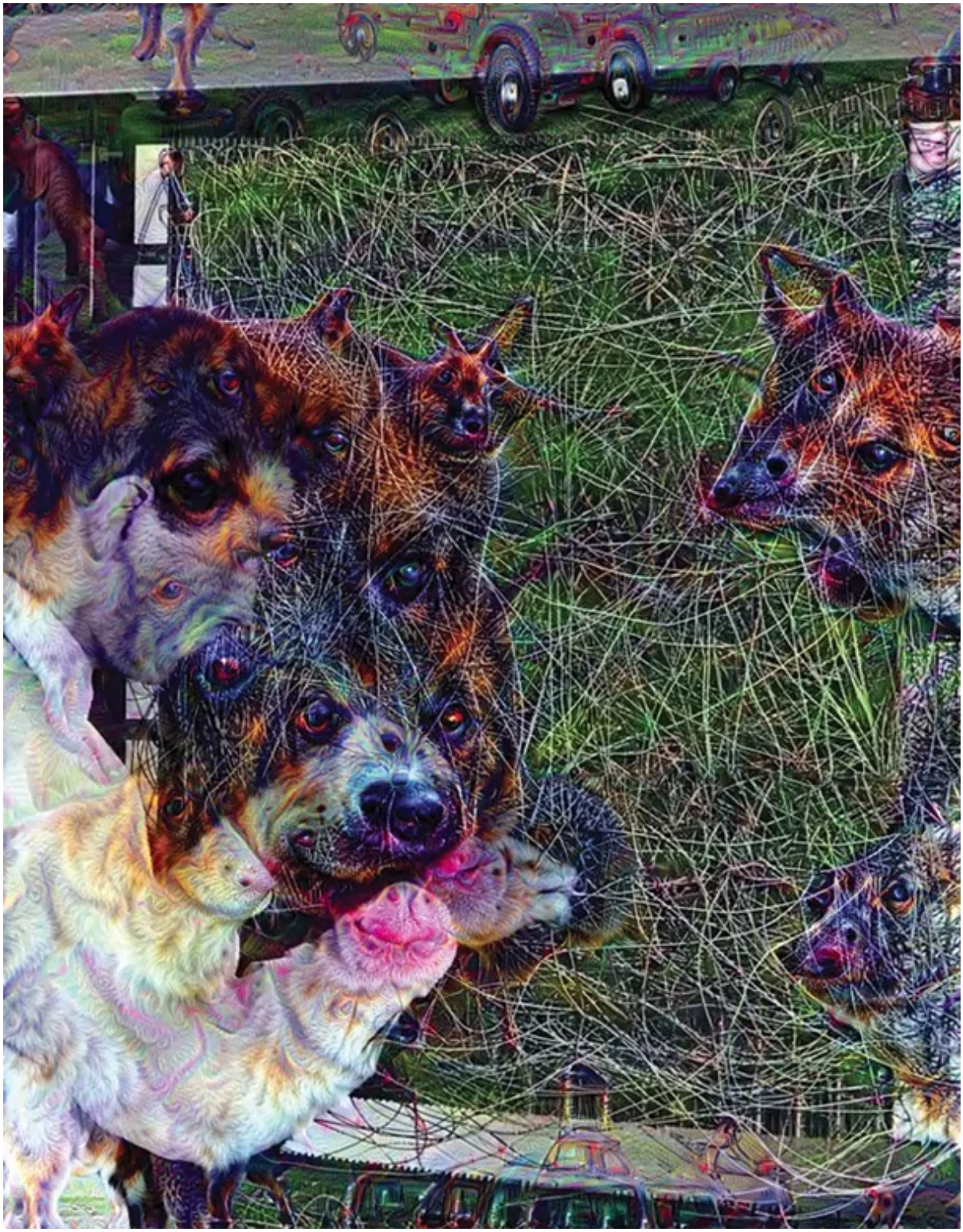
After she finished cancer treatment last year, Barzilay and her students began working with doctors at Massachusetts General Hospital to develop a system capable of mining pathology reports to identify patients with specific clinical characteristics that researchers might want to study. However, Barzilay understood that the system would need to explain its reasoning. So, together with Jaakkola and a student, she added a step: the system extracts and highlights snippets of text that are representative of a pattern it has discovered. Barzilay and her students are also developing a deep-learning algorithm capable of finding early signs of breast cancer in mammogram images, and they aim to give this system some ability to explain its reasoning, too. "You really need to have a loop where the machine and the human collaborate," -Barzilay says.

The U.S. military is pouring billions into projects that will use machine learning to pilot vehicles and aircraft, identify targets, and help analysts sift through huge piles of intelligence data. Here more than anywhere else, even

more than in medicine, there is little room for algorithmic mystery, and the Department of Defense has identified explainability as a key stumbling block.

David Gunning, a program manager at the Defense Advanced Research Projects Agency, is overseeing the aptly named Explainable Artificial Intelligence program. A silver-haired veteran of the agency who previously oversaw the DARPA project that eventually led to the creation of Siri, Gunning says automation is creeping into countless areas of the military. Intelligence analysts are testing machine learning as a way of identifying patterns in vast amounts of surveillance data. Many autonomous ground vehicles and aircraft are being developed and tested. But soldiers probably won't feel comfortable in a robotic tank that doesn't explain itself to them, and analysts will be reluctant to act on information without some reasoning. "It's often the nature of these machine-learning systems that they produce a lot of false alarms, so an intel analyst really needs extra help to understand why a recommendation was made," Gunning says.

This March, DARPA chose 13 projects from academia and industry for funding under Gunning's program. Some of them could build on work led by Carlos Guestrin, a professor at the University of Washington. He and his colleagues have developed a way for machine-learning systems to provide a rationale for their outputs. Essentially, under this method a computer automatically finds a few examples from a data set and serves them up in a short explanation. A system designed to classify an e-mail message as coming from a terrorist, for example, might use many millions of messages in its training and decision-making. But using the Washington team's approach, it could highlight certain keywords found in a message. Guestrin's group has also devised ways for image recognition systems to hint at their reasoning by highlighting the parts of an image that were most significant.



ADAM FERRISS

One drawback to this approach and others like it, such as Barzilay's, is that the explanations provided will always be simplified, meaning some vital information may be lost along the way. "We haven't achieved the whole dream, which is where AI has a conversation with you, and it is able to explain," says Guestrin. "We're a long way from having truly interpretable AI."

It doesn't have to be a high-stakes situation like cancer diagnosis or military maneuvers for this to become an issue. Knowing AI's reasoning is also going to be crucial if the technology is to become a common and useful part of our daily lives. Tom Gruber, who leads the Siri team at Apple, says explainability is a key consideration for his team as it tries to make Siri a smarter and more capable virtual assistant. Gruber wouldn't discuss specific plans for Siri's future, but it's easy to imagine that if you receive a restaurant recommendation from Siri, you'll want to know what the reasoning was. Ruslan Salakhutdinov, director of AI research at Apple and an associate professor at Carnegie Mellon University, sees explainability as the core of the evolving relationship between humans and intelligent machines. "It's going to introduce trust," he says.

Read Next

AI's Language Problem
Machines that truly understand language would be incredibly useful. But we don't know how to build them.

Just as many aspects of human behavior are impossible to explain in detail, perhaps it won't be possible for AI to explain everything it does. "Even if somebody can give you a reasonable-sounding explanation [for his or her actions], it probably is incomplete, and the same could very well be true for AI," says Clune, of the University of Wyoming. "It might just be part of the nature of intelligence that only part of it is exposed to rational explanation. Some of it is just instinctual, or subconscious, or inscrutable."

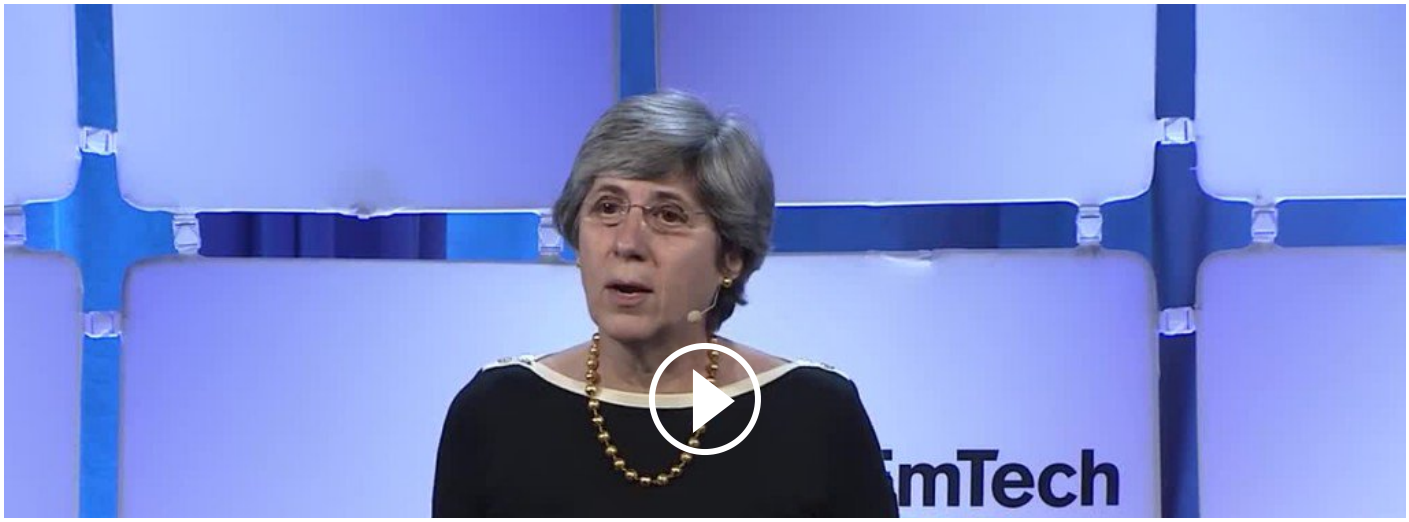
If that's so, then at some stage we may have to simply trust AI's judgment or do without using it. Likewise, that judgment will have to incorporate social intelligence. Just as society is built upon a contract of expected behavior, we will need to design AI systems to respect and fit with our social norms. If we are to create robot tanks and other killing machines, it is important that their decision-making be consistent with our ethical judgments.

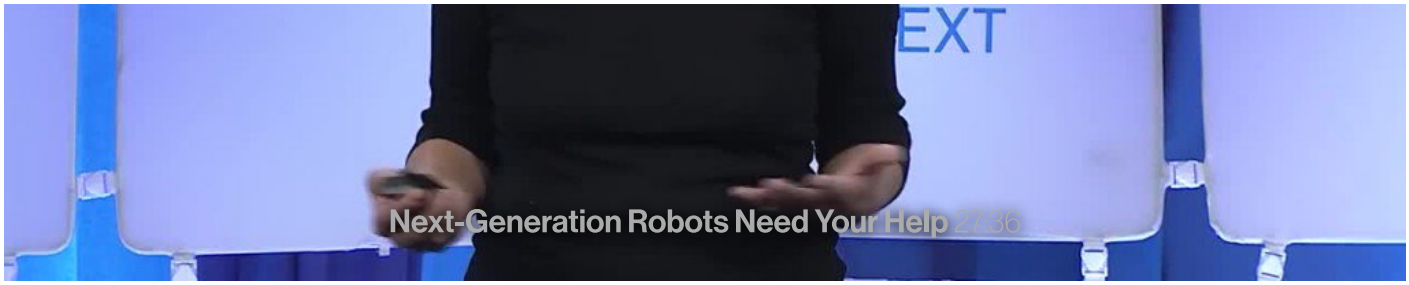
To probe these metaphysical concepts, I went to Tufts University to meet with Daniel Dennett, a renowned philosopher and cognitive scientist who studies consciousness and the mind. A chapter of Dennett's latest book, *From Bacteria to Bach and Back*, an encyclopedic treatise on consciousness, suggests that a natural part of the evolution of intelligence itself is the creation of systems capable of performing tasks their creators do not know how to do. "The question is, what accommodations do we have to make to do this wisely—what standards do we demand of them, and of ourselves?" he tells me in his cluttered office on the university's idyllic campus.

He also has a word of warning about the quest for explainability. "I think by all means if we're going to use these things and rely on them, then let's get as firm a grip on how and why they're giving us the answers as possible," he says. But since there may be no perfect answer, we should be as cautious of AI explanations as we are of each other's—no matter how clever a machine seems. "If it can't do better than us at explaining what it's doing," he says, "then don't trust it."

Related Video

More videos





Recommended for You

01 The smartphone app that can tell you're depressed before you know it yourself

02 MIT has just announced a $1 billion plan to create a new college for AI

03 We can now customize cancer cures, tumor by tumor

04 Why we can't quit the QWERTY keyboard

05 The 8-dimensional space that must be searched for alien life

More from Intelligent Machines

Artificial intelligence and robots are transforming how we work and live.

01 **Your next doctor's appointment might be with an AI**

A new wave of chatbots are replacing physicians and providing frontline medical advice—but are they as good as the real thing?

by Douglas Heaven

02 **Why we can't quit the QWERTY keyboard**

We've been using it to type for 144 years. Here's why it works, and what it would take for us to give it up.

by Rachel Metz

03 **Neural networks don't understand what optical illusions are**

Machine-vision systems can match humans at recognizing faces and can even create realistic synthetic faces. But researchers have discovered that the same systems cannot recognize optical illusions, which means they also can't create new ones.

by Emerging Technology from the arXiv

More from Intelligent Machines

Subscribe

See details+

---

*Prices are for U.S. residents only
See international prices