

KEITH RANKIN

Intelligent Machines

The Dark Secret at the Heart of AI

CONTENT COLLECTION

No one really knows how the most advanced algorithm problem.

by Will Knight April 11, 2017

**Wireless Technology Innovations
Lead the Way to a Smartly
Connected Future**

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 no reason to think that the techniques will be able to diagnose deadly diseases, make trading decisions, and do countless other things that will transform our industries.

CONTENT COLLECTION

**Wireless Technology Innovations
Lead the Way to a Smartly
Connected Future**

But this won't happen—or shouldn't happen—unless we're 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 e

these apps cannot fully explain their behavior.

CONTENT COLLECTION

**Wireless Technology Innovations
Lead the Way to a Smartly
Connected Future**

This raises mind-boggling questions. As the tech might soon cross some threshold beyond which u of faith. Sure, we humans can't always truly expla 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.

CONTENT COLLECTION

**Wireless Technology Innovations
Lead the Way to a Smartly
Connected Future**



CONTENT COLLECTION

Wireless Technology Innovations Lead the Way to a Smartly Connected Future

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."

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 prediction, to reassure them that it is accurate and change in the drugs someone is being prescribed. models," Dudley says ruefully, "but we don't know

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

CONTENT COLLECTION

**Wireless Technology Innovations
Lead the Way to a Smartly
Connected Future**

Artificial intelligence hasn't always been this way. From the outset, there were two schools of thought regarding how understandable, or

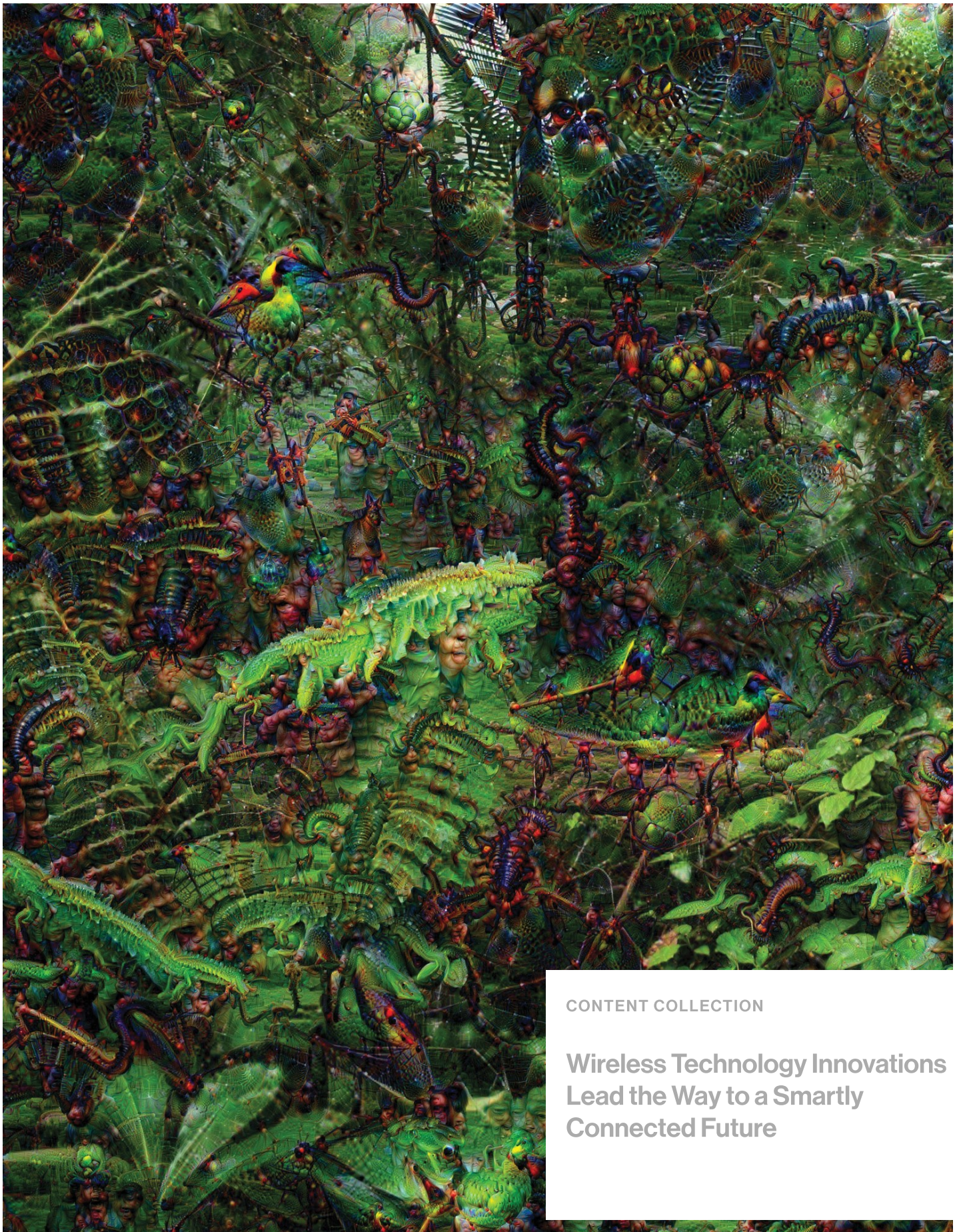
explainable, AI ought to be. Many thought it made the most sense to build machines that reasoned according to rules and logic, making their inner workings transparent to anyone who cared to examine some code. Others felt that intelligence would more easily emerge if machines took inspiration from biology, 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 the machine by hand. Deep learning has transformed machine translation and dramatically improved machine translation. It can guide all sorts of key decisions in medicine, finance, and beyond.

CONTENT COLLECTION

Wireless Technology Innovations
Lead the Way to a Smartly
Connected Future



CONTENT COLLECTION

Wireless Technology Innovations
Lead the Way to a Smartly
Connected Future

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.

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

CONTENT COLLECTION

Wireless Technology Innovations
Lead the Way to a Smartly
Connected Future

“It
of
intelligence that

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.

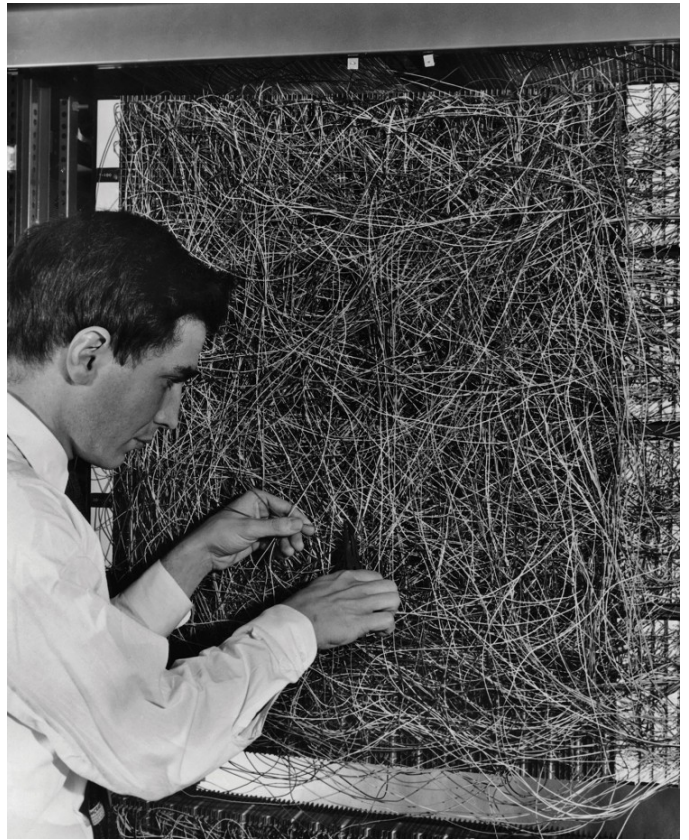
Further progress has been made using ideas borrowed from neuroscience and cognitive science. A team led by Jeff Clune, an assistant professor at the University of Wyoming, has employed the AI equivalent of optical illusions to test deep neural networks. In 2015, Clune's group showed how certain images could fool the system into perceiving things that aren't there, because the images contained low-level patterns the system searches for. One of Clune's images, a flamingo, was perceived as a school bus.

Jason Yosinski, also built a tool that acts like a proctologist. His tool targets any neuron in the middle of the network and finds for the image that activates it the most. The images are often abstract (imagine an impressionistic take on a flamingo or a school bus), highlighting the mysterious nature of the machine's perceptual abilities.

only part of it is exposed to rational explanation. Some of it is just instinctual."

CONTENT COLLECTION

**Wireless Technology Innovations
Lead the Way to a Smartly
Connected Future**



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

CONTENT COLLECTION

**Wireless Technology Innovations
Lead the Way to a Smartly
Connected Future**

CONTENT COLLECTION

Ferriss was inspired to run Cornell's artificial neural network through Dee above and below.

ADAM FERRISS

**Wireless Technology Innovations
Lead the Way to a Smartly
Connected Future**

We need more than a glimpse of AI's thinking, however. It is the interplay of calculations in a 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.”

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 snippet representative of a pattern it has discovered. Barz are also developing a deep-learning algorithm can signs of breast cancer in mammogram images, an system some ability to explain its reasoning, too. “You really need to

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

CONTENT COLLECTION

Wireless Technology Innovations
Lead the Way to a Smartly
Connected Future

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 real understanding of the nature of these machine-learning systems that trigger the alarms, so an intel analyst really needs extra help when a recommendation was made,” Gunning says.

This March, DARPA chose 13 projects from academic 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

Recommended for You

- 01 We Need More Alternatives to Facebook
- 02 FDA Opens Genetic Floodgates with 23andMe Decision
- 03 Meet the Guys Who Sold “Neuralink” to Elon Musk without Even Realizing It
- 04 Can This Dumb Phone Free Us from Smartphone Addiction?
- 05 Can “Digital Therapeutics” Be as Good as Drugs?

CONTENT COLLECTION

Wireless Technology Innovations Lead the Way to a Smartly Connected Future

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.

CONTENT COLLECTION

**Wireless Technology Innovations
Lead the Way to a Smartly
Connected Future**



CONTENT COLLECTION

ADAM FERRISS

One drawback to this approach and others like it, that the explanations provided will always be sim
vital information may be lost along the way. “We l
whole dream, which is where AI has a conversation with you, and it is

**Wireless Technology Innovations
Lead the Way to a Smartly
Connected Future**

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 of his or her actions[, it probably is incomplete, and that will be true for AI,” says Clune, of the University of Cambridge. “It will just be part of the nature of intelligence that only a partial, not a rational explanation. Some of it is just instinctual and inscrutable.”

CONTENT COLLECTION

Wireless Technology Innovations
Lead the Way to a Smartly
Connected Future

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."

CONTENT COLLECTION

Wireless Technology Innovations
Lead the Way to a Smartly
Connected Future

**Weren't able to make it to EmTech D
We've got you covered.**

Watch videos here

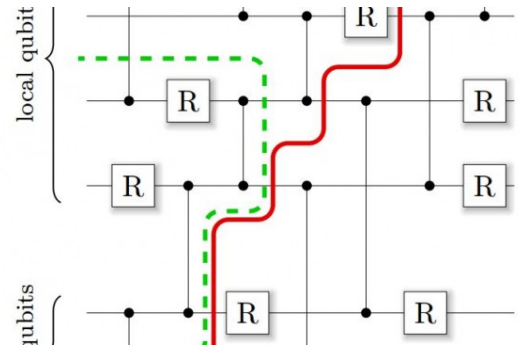
More from Intelligent Machines

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

01 A Milestone for Quantum Computing

To find out whether quantum computers will work properly, scientists must simulate them on a classical computer. Now a record-breaking experiment has simulated the largest quantum computer yet.

by Emerging Technology from the arXiv



02 This Robot Will Swing Over Crops like Tarzan

Researchers actually modeled this plant-watching robot on sloths.

by Jamie Condliffe



03 Robotics, Smart Materials, and their Future Impact for Humans

We are on the cusp of a robotics revolution in which the boundaries between artificial intelligence and biology are blurring.

by Jonathan Rossiter



More from Intelligent Machines

CONTENT COLLECTION

**Wireless Technology Innovations
Lead the Way to a Smartly
Connected Future**

Want more award-winning jc
and become an Insider.

Insider Premium \$179.95/year*

Our award winning magazine, unlimited access to our story archive, special discounts to MIT Technology Review Events, and exclusive content.

Subscribe



See details+

Insider Plus \$79.95/year* BEST VALUE

Everything included in Insider Basic, plus ad-free web experience, select discounts to partner offerings and MIT Technology Review events

Subscribe



See details+

Insider Basic \$29.95/year*

Six issues of our award winning magazine and daily delivery of The Download, our newsletter of what's important in technology and innovation.

Subscribe



See details+

CONTENT COLLECTION

**Wireless Technology Innovations
Lead the Way to a Smartly
Connected Future**

*Prices are for U.S. residents only
[See international prices](#)

