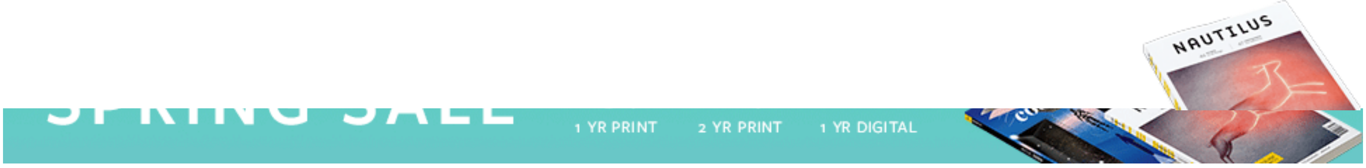
NAUTILUS



Purchase This Artwork

NUMBERS | ARTIFICIAL INTELLIGENCE

Is Artificial Intelligence Permanently Inscrutable?

Despite new biology-like tools, some insist interpretation is impossible.

BY AARON M. BORNSTEIN
ILLUSTRATION BY EMMANUEL POLANCO
SEPTEMBER 1, 2016

ADD A COMMENT FACEBOOK TWITTER EMAIL SHARING

Dmitry Malioutov can't say much about what he built.

As a research scientist at IBM, Malioutov spends part of his time building machine learning systems that solve difficult problems faced by IBM's corporate clients. One such program was meant for a large insurance corporation. It was a challenging assignment, requiring a sophisticated algorithm. When it came time to describe the results to his client, though, there was a wrinkle. "We couldn't explain the model to them because they didn't have the training in machine learning."

In fact, it may not have helped even if they were machine learning experts. That's because the model was an artificial neural network, a program that takes in a given type of data—in this case, the insurance company's customer records—and finds patterns in them. These networks have been in practical use for over half a century, but lately they've seen a resurgence, powering breakthroughs in everything from speech recognition and language translation to Go-playing robots and self-driving cars.

HIDDEN MEANINGS: In neural networks, data is passed from layer to layer, undergoing simple transformations at each step. Between the input and output layers are hidden layers, groups of nodes and connections that often bear no human-interpretable patterns or obvious connections to either input or output. "Deep" networks are those with many hidden layers.

Michael Nielsen / NeuralNetworksandDeepLearning.com

As exciting as their performance gains have been, though, there's a troubling fact about modern neural networks: Nobody knows quite how they work. And that means no one can predict when they might fail.

Take, for example, an episode recently reported by machine learning researcher Rich Caruana and his colleagues. They described the experiences of a team at the University of Pittsburgh Medical Center who were using machine learning to predict whether pneumonia patients might develop severe complications. The goal was to send patients at low risk for complications to outpatient treatment, preserving hospital beds and the attention of medical staff. The team tried several different methods, including various kinds of neural networks, as well as software-generated decision trees that produced clear, human-readable rules.

The neural networks were right more often than any of the other methods. But when the researchers and doctors took a look at the human-readable rules, they noticed something disturbing: One of the rules instructed doctors to send home pneumonia patients who already had asthma, despite the fact that asthma sufferers are known to be extremely vulnerable to complications.

The model did what it was told to do: Discover a true pattern in the data. The poor advice it produced was the result of a quirk in that data. It was hospital policy to send asthma sufferers with pneumonia to intensive care, and this policy worked so well that asthma sufferers almost never developed severe complications. Without the extra care that had shaped the hospital's patient records, outcomes could have been dramatically different.

ALSO IN ARTIFICIAL INTELLIGENCE

The Unexpected Humanity of Robot Soccer

By Seth Frey & Patrick House

When Google's AlphaGo computer program triumphed over a Go expert earlier this year, a human member of the Google team had to physically move the pieces. Manuela Veloso, the head of Carnegie Mellon's machine learning department, would have done it...**READ MORE**

The hospital anecdote makes clear the practical value of interpretability. "If the rule-based system had learned that asthma lowers risk, certainly the neural nets had learned it, too," wrote Caruana and colleagues—but the neural net wasn't human-interpretable, and its bizarre conclusions about asthma patients might have been difficult to diagnose.[1] If there hadn't been an interpretable model, Malioutov cautions, "you could accidentally kill people."

This is why so many are reluctant to gamble on the mysteries of neural networks. When Malioutov presented his accurate but inscrutable neural network model to his own corporate client, he also offered them an alternative, rule-based model whose workings he could communicate in simple terms. This second, interpretable, model was less accurate than the first, but the client decided to use it anyway—despite being a mathematically sophisticated insurance company for which every percentage point of accuracy mattered. "They could relate to [it] more," Malioutov says. "They really value intuition highly."

Even governments are starting to show concern about the increasing influence of inscrutable neural-network oracles. The European Union recently proposed to establish a "right to explanation," which allows citizens to demand transparency for algorithmic decisions.[2] The legislation may be difficult to implement, however, because the legislators didn't specify exactly what "transparency" means. It's unclear whether this omission stemmed from ignorance of the problem, or an appreciation of its complexity.

Some researchers hope to eliminate the need to choose—to let us have our many-layered cake, and understand it, too.

In fact, some believe that such a definition might be impossible. At the moment, though we can know everything there is to know about what neural networks are doing—they are, after all, just computer programs—we can discern very little about how or why they are doing it. The networks are made up of many, sometimes millions, of individual units, called neurons. Each neuron converts many numerical inputs into a single numerical output, which is then passed on to one or more other neurons. As in brains, these neurons are divided into "layers," groups of cells that take input from the layer below and send their output to the layer above.

Neural networks are trained by feeding in data, then adjusting the connections between layers until the network's calculated output matches the known output (which usually consists of categories) as closely as possible. The incredible results of the past few years are thanks to a series of new techniques that make it possible to quickly train deep networks, with many layers between the first input and the final output. One popular deep network called AlexNet is used to categorize photographs—labeling them according to such fine distinctions as whether they contain a Shih Tzu or a Pomeranian. It consists of over 60 million "weights," each of which tell each neuron how much attention to pay to each of its inputs. "In order to say you have some understanding of the network," says Jason Yosinski, a computer scientist affiliated with Cornell University and Geometric Intelligence, "you'd have to have some understanding of these 60 million numbers."

Even if it were possible to impose this kind of interpretability, it may not always be desirable. The requirement for interpretability can be seen as another set of constraints, preventing a model from a "pure" solution that pays attention only to the input and output data it is given, and potentially reducing accuracy. At a DARPA conference early this year, program manager David Gunning summarized the trade-off in a chart that shows deep networks as the least understandable of modern methods. At the other end of the spectrum are decision trees, rule-based systems that tend to prize explanation over efficacy.

WHAT VS. WHY: Modern learning algorithms show a tradeoff between human interpretability, or explainability, and their accuracy. Deep learning is both the most accurate and the least interpretable.

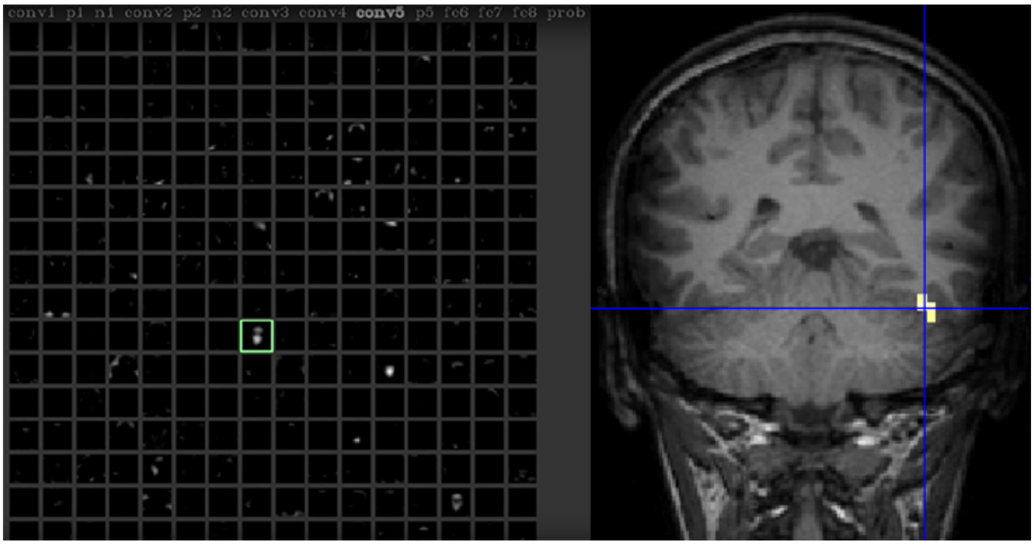
Darpa

The result is that modern machine learning offers a choice among oracles: Would we like to know *what* will happen with high accuracy, or *why* something will happen, at the expense of accuracy? The "why" helps us strategize, adapt, and know when our model is about to break. The "what" helps us act appropriately in the immediate future.

It can be a difficult choice to make. But some researchers hope to eliminate the need to choose—to allow us to have our many-layered cake, and understand it, too. Surprisingly, some of the most promising avenues of research treat neural networks as experimental objects—after the fashion of the biological science that inspired them to begin with—rather than analytical, purely mathematical objects. Yosinski, for example, says he is trying to understand deep networks "in the way we understand animals, or maybe even humans." He and other computer scientists are importing techniques from biological research that peer inside networks after the fashion of neuroscientists peering into brains: probing individual components, cataloguing how their internals respond to small changes in inputs, and even removing pieces to see how others compensate.

Having built a new intelligence from scratch, scientists are now taking it apart, applying to these virtual organisms the digital equivalents of a microscope and scalpel.

Yosinski sits at a computer terminal, talking into a webcam. The data from the webcam is fed into a deep neural net, while the net itself is being analyzed, in real time, using a software toolkit Yosinski and his colleagues developed called the Deep Visualization toolkit. Clicking through several screens, Yosinski zooms in on one neuron in the network. "This neuron seems to respond to faces," he says in a video record of the interaction.[3] Human brains are also known to have such neurons, many of them clustered in a region of the brain called the fusiform face area. This region, discovered over the course of multiple studies beginning in 1992,[4, 5] has become one of the most reliable observations in human neuroscience. But where those studies required advanced techniques like positron emission tomography, Yosinski can peer at his artificial neurons through code alone.


conv1 p1 n1 conv2 p2 n2 conv3 conv4 conv5 p5 fc6 fc7 fc8 prob

BRAIN ACTIVITY: A single neuron in a deep neural net (highlighted by a green box) responds to Yosinski's face, just as a distinct part of the human brain reliably responds to faces (highlighted in yellow).

Left: Jason Yosinski, *et al.* Understanding Neural Networks Through Deep Visualization. Deep Learning Workshop, International Conference on Machine Learning (ICML) (2015). Right: Maximilian Riesenhuber, Georgetown University Medical Center

This approach lets him map certain artificial neurons to human-understandable ideas or objects, like faces, which could help turn neural networks into intuitive tools. His program can also highlight which aspects of a picture are most important to stimulating the face neuron. "We can see that it would respond even more strongly if we had darker eyes, and rosier lips," he says.

To Cynthia Rudin, professor of computer science and electrical and computer engineering at Duke University, these "post-hoc" interpretations are by nature problematic. Her research focuses on building rule-based machine learning systems applied to domains like prison sentencing and medical diagnosis, where human-readable interpretations are possible—and critically important. But for problems in areas like vision, she says, "Interpretations are completely within the eye of the beholder." We can simplify a network response by identifying a face neuron, but how can we be certain that's really what it's looking for? Rudin's concerns echo the famous dictum that there may be no simpler model of the visual system than the visual system itself. "You could have many explanations for what a complex model is doing," she says. "Do you just pick the one you 'want' to be correct?"

Yosinski's toolkit can, in part, counter these concerns by working backward, discovering what the network itself "wants" to be correct—a kind of artificial ideal. The program starts with raw static, then adjusts it, pixel by pixel, tinkering with the image using the reverse of the process that trained the network. Eventually it finds a picture that elicits the maximum possible response of a given neuron. When this method is applied to AlexNet neurons, it produces caricatures that, while ghostly, unquestionably evoke the labeled categories.

IDEALIZED CATS: Examples of synthetic ideal cat faces, generated by the Deep Visualization toolkit. These faces are generated by tweaking a generic starting image pixel by pixel, until a maximum response from AlexNet's face neuron is achieved.

Jason Yosinski, *et al.* Understanding Neural Networks Through Deep Visualization. Deep Learning Workshop, International Conference on Machine Learning (ICML) (2015).

This seems to support his claim that the face neurons are indeed looking for faces, in some very general sense. But there's a catch. To generate those pictures, Yosinski's procedure relies on a statistical constraint (called a natural image prior) that confines it to producing images that match the sorts of structure that one finds in pictures of real-world objects. When he removes those rules, the toolkit still settles on an image that it labels with maximum confidence, but that image is pure static. In fact, Yosinski has shown that in many cases, the majority of images that AlexNet neurons prefer appear to humans as static. He readily admits that "it's pretty easy to figure out how to make the networks say something extreme."

To avoid some of these pitfalls, Dhruv Batra, an assistant professor of electrical and computer engineering at Virginia Tech, takes a higher-level experimental approach to interpreting deep networks. Rather than trying to find patterns in their internal structure —"people smarter than me have worked on it," he demurs—he probes how the networks behave using, essentially, a robotic version of eye tracking. His group, in a project led by graduate students Abhishek Das and Harsh Agrawal, asks a deep net questions about an image, like whether there are drapes on a window in a given picture of a room.[6] Unlike AlexNet or similar systems, Das' network is designed to focus on only a small patch of an image at a time. It moves its virtual eyes around the picture until it has decided it has enough information to answer the question. After sufficient training, the deep network performs extremely well, answering about as accurately as the best humans.

Trained machines are exquisitely well suited to their environment—and ill-adapted to any other.

Das, Batra, and their colleagues then try to get a sense of how the network makes its decisions by investigating where in the pictures it chooses to look. What they have found surprises them: When answering the question about drapes, the network doesn't even bother looking for a window. Instead, it first looks to the bottom of the image, and stops looking if it finds a bed. It seems that, in the dataset used to train this neural net, windows with drapes may be found in bedrooms.

While this approach does reveal some of the inner workings of the deep net, it also reinforces the challenge presented by interpretability. "What machines are picking up on are not facts about the world," Batra says. "They're facts about the dataset." That the machines are so tightly tuned to the data they are fed makes it difficult to extract general rules about how they work. More importantly, he cautions, if you don't know how it works, you don't know how it will fail. And when they do they fail, in Batra's experience, "they fail spectacularly disgracefully."

Some of the obstacles faced by researchers like Yosinski and Batra will be familiar to scientists studying the human brain. Questions over the interpretations of neuroimaging, for example, are common to this day, if not commonly heeded. In a 2014 review of the field, cognitive neuroscientist Martha Farah wrote that "the worry ... is that [functional brain] images are more researcher inventions than researcher observations."[7] The appearance of these issues in very different realizations of intelligent systems suggests that they could be obstacles, not to the study of this or that kind of brain, but to the study of intelligence itself.

Is chasing interpretability a fool's errand? In a 2015 blog post entitled "The Myth of Model Interpretability," Zachary Lipton, of the University of California, San Diego, offered a critical perspective on both the motivations behind interpreting neural networks, and the value of building interpretable machine learning models for huge datasets in the first place. He submitted a provocative paper on this subject to a workshop (organized by Malioutov and two of his colleagues) on Human Interpretability at this year's International Conference on Machine Learning (ICML).[8]

Lipton points out that many scholars disagree over the very concept of interpretability, which suggests to him either that interpretability is poorly understood—or that there are many equally valid meanings. In either case, chasing interpretability may not satisfy our desire for a straightforward, plain-English description of a neural net output. In his blog post, Lipton argued that, when it comes to enormous datasets, researchers have the option to resist the impulse to interpret and could, instead, "place faith in empirical success." One purpose of the field, he argued, is to "build models which can learn from a greater number of features than any human could consciously account for," and interpretability could keep such models from reaching their full potential.

But this ability is both feature and failing: If we don't understand how network output is generated, then we can't know what aspects of the input were necessary, or even what might be considered input at all. Case in point: In 1996, Adrian Thompson of Sussex University used software to design a circuit by applying techniques similar to those that train deep networks today. The circuit was to perform a straightforward task: discriminate between two audio tones. After thousands of iterations, shuffling and rearranging circuit components, the software found a configuration that performed the task nearly perfectly.

Thompson was surprised, however, to discover that the circuit used fewer components than any human engineer would have used—including several that were not physically connected to the rest, and yet were somehow still necessary for the circuit to work properly.

He took to dissecting the circuit. After several experiments, he learned that its success exploited subtle electromagnetic interference between adjacent components. The disconnected elements influenced the circuit by causing small fluctuations in local electrical fields. Human engineers usually guard against these interactions, because they are unpredictable. Sure enough, when Thompson copied the same circuit layout to another batch of components—or even changed the ambient temperature—it failed completely.

The circuit exhibited a hallmark feature of trained machines: They are as compact and simplified as they can be, exquisitely well suited to their environment—and ill-adapted to any other. They pick up on patterns invisible to their engineers; but can't know which of those patterns exist nowhere else. Machine learning researchers go to great lengths to avoid this phenomenon, called "overfitting," but as these algorithms are used in more and more dynamic situations, their brittleness will inevitably be exposed.

To Sanjeev Arora, a professor of computer science at Princeton University, this fact is the primary motivation to seek interpretable models that allow humans to intervene and adjust the networks. Arora points to two problems that could represent hard limits on the capabilities of machines in the absence of interpretability. One is "composability"—when the task at hand involves many different decisions (as with Go, or self-driving cars), networks can't efficiently learn which are responsible for a failure. "Usually when we design things, we understand the different components and then we put them together," he says. This allows humans to adjust components that aren't appropriate for a given environment.

The other problem with leaving interpretability unsolved is what Arora calls "domain adaptability"—the ability to flexibly apply knowledge learned in one setting to another. This is a task human learners do very well, but machines can fail at in surprising ways. Arora describes how programs can be catastrophically incapable of adjusting to even subtle contextual shifts of the sort that humans handle with ease. For instance, a network trained to parse human language by reading formal documents, like Wikipedia, can fail completely in more vernacular settings, like Twitter.

By this view, interpretability seems essential. But do we understand what we mean by the word? Pioneering computer scientist Marvin Minsky coined the phrase "suitcase word" to describe many of the terms—such as "consciousness" or "emotion"—we use when we talk about our own intelligence.[9] These words, he proposed, reflect the workings of many different underlying processes, which are locked inside the "suitcase." As long as we keep investigating these words as stand-ins for the more fundamental concepts, the argument went, our insight will be limited by our language. In the study of intelligence, could interpretability itself be such a suitcase word?

While many of the researchers I spoke to are optimistic that theorists will someday unlock the suitcase and discover a single, unifying set of principles or laws that govern machine (and, perhaps, human) learning, akin to Newton's *Principia*, others warn that there is little reason to expect this. Massimo Pigliucci, a professor of philosophy at City University of New York, cautions that "understanding" in the natural sciences—and, by extension, in artificial intelligence—might be what Ludwig Wittgenstein, anticipating Minsky, called a "cluster concept," one which could admit many, partly distinct, definitions. If "understanding" in this field does come, he says, it could be of the sort found not in physics, but evolutionary biology. Rather than *Principia*, he says, we might expect *Origin of the Species*.

This doesn't mean, of course, that deep networks are a harbinger of some new kind of autonomous life. But they could turn out to be just as difficult to understand as life. The field's incremental, experimental approaches and post-hoc interpretations may not be some kind of desperate feeling around in the dark, hoping for theory to shine a light. They could, instead, be the only kind of light we can expect. Interpretability may emerge piecemeal, as a set of prototypic examples of "species" arranged in a taxonomy defined by inference and contingent, context-specific explanations.

At the ICML workshop's close, some of the presenters appeared on a panel to try and define "interpretability." There were as many responses as there were panelists. After some discussion the group seemed to find consensus that "simplicity" was necessary for a model to be interpretable. But, when pressed to define simplicity, the group again diverged. Is the "simplest" model the one that relies on the fewest number of features? The one that makes the sharpest distinctions? Is it the smallest program? The workshop closed without an agreed-upon answer, leaving the definition of one inchoate concept replaced by another.

As Malioutov puts it, "Simplicity is not so simple."

Aaron M. Bornstein is a researcher at the Princeton Neuroscience Institute. His research investigates how we use memories to make sense of the present and plan for the future.

References

1. Caruana, R., *et. al* Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 1721-1730 (2015).

2. Metz, C. Artificial Intelligence Is Setting Up the Internet for a Huge Clash with Europe. Wired.com (2016).

3. Yosinski, J., Clune, J., Nguyen, A., Fuchs, T., & Lipson, H. Understanding neural networks through deep visualization. arXiv:1506.06579 (2015).

4. Sergent, J., Ohta, S., & MacDonald, B. Functional neuroanatomy of face and object processing. A positron emission tomography study. *Brain* **115**, 15–36 (1992).

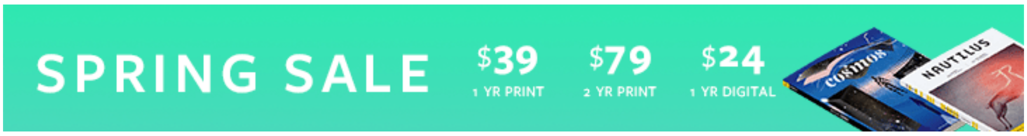
5. Kanwisher. N., McDermott, J., & Chun, M.M. The fusiform face area: A module in human extrastriate cortex specialized for face perception. *The Journal of Neuroscience* **17**, 4302–4311 (1997).

6. Das, A., Agrawal, H., Zitnick, C.L., Parikh, D., & Batra, D. Human attention in visual question answering: Do humans and deep networks look at the same regions? Conference on Empirical Methods in Natural Language Processing (2016).

7. Farah, M.J. Brain images, babies, and bathwater: Critiquing critiques of functional neuroimaging. *Interpreting Neuroimages: An Introduction to the Technology and Its Limits* **45**, S19-S30 (2014).

8. Lipton, Z.C. The mythos of model interpretability. arXiv:1606.03490 (2016).

9. Brockman, J. Consciousness Is a Big Suitcase: A talk with Marvin Minsky. Edge.org (1998).



JOIN THE DISCUSSION

NEXT ARTICLE:

RELATED ARTICLES:

IDEAS
How Much More Can We Learn About the Universe?
By Lawrence M. Krauss

NUMBERS
Teaching Me Softly
By Alan S. Brown

NUMBERS
A.I. Has Grown Up and Left Home
By David Auerbach

NUMBERS
Don't Worry, Smart Machines Will Take Us With Them
By Stephen Hsu

ABOUT
CONTACT / WORK WITH US
FAQ
PRIME
SUBSCRIBE

AWARDS AND PRESS
DONATE
MEDIA KIT
RSS
TERMS OF SERVICES

NAUTILUS: SCIENCE CONNECTED

Nautilus is a different kind of science magazine. We deliver big-picture science by reporting on a single monthly topic from multiple perspectives. Read a new chapter in the story every Thursday.

Matter, Biology, Numbers, Ideas, Culture, *Connected*