Can A.I. Be Taught to Explain Itself?

Cliff Kuang

In September, Michal Kosinski published a study that he feared might end his career. The Economist broke the news first, giving it a self-consciously anodyne title: “Advances in A.I. Are Used to Spot Signs of Sexuality.” But the headlines quickly grew more alarmed. By the next day, the Human Rights Campaign and Glaad, formerly known as the Gay and Lesbian Alliance Against Defamation, had labeled Kosinski’s work “dangerous” and “junk science.” (They claimed it had not been peer reviewed, though it had.) In the next week, the tech-news site The Verge had run an article that, while carefully reported, was nonetheless topped with a scorching headline: “The Invention of A.I. ‘Gaydar’ Could Be the Start of Something Much Worse.”

Kosinski has made a career of warning others about the uses and potential abuses of data. Four years ago, he was pursuing a Ph.D. in psychology, hoping to create better tests for signature personality traits like introversion or openness to change. But he and a collaborator soon realized that Facebook might render personality tests superfluous: Instead of asking if someone liked poetry, you could just see if they “liked” Poetry Magazine. In 2014, they published a study showing that if given 200 of a user’s likes, they could predict that person’s personality-test answers better than their own romantic partner could.

After getting his Ph.D., Kosinski landed a teaching position at the Stanford Graduate School of Business and soon started looking for new data sets to investigate. One in particular stood out: faces. For decades, psychologists have been leery about associating personality traits with physical characteristics, because of the lasting taint of phrenology and eugenics; studying faces this way was, in essence, a taboo. But to understand what that taboo might reveal when questioned, Kosinski knew he couldn’t rely on a human judgment.

Kosinski first mined 200,000 publicly posted dating profiles, complete with pictures and information ranging from personality to political views. Then he poured that data into an open-source facial-recognition algorithm — a so-called deep neural network, built by researchers at Oxford University — and asked it to find correlations between people's faces and the information in their profiles. The algorithm failed to turn up much, until, on a lark, Kosinski turned its attention to sexual orientation. The results almost defied belief. In previous research, the best any human had done at guessing sexual orientation from a profile picture was about 60 percent — slightly better than a coin flip. Given five pictures of a man, the deep neural net could predict his sexuality with as much as 91 percent accuracy. For women, that figure was lower but still remarkable: 83 percent.

Much like his earlier work, Kosinski's findings raised questions about privacy and the potential for discrimination in the digital age, suggesting scenarios in which better programs and data sets might be able to deduce anything from political leanings to criminality. But there was another question at the heart of Kosinski's paper, a genuine mystery that went almost ignored amid all the media response: *How* was the computer doing what it did? What was it seeing that humans could not?

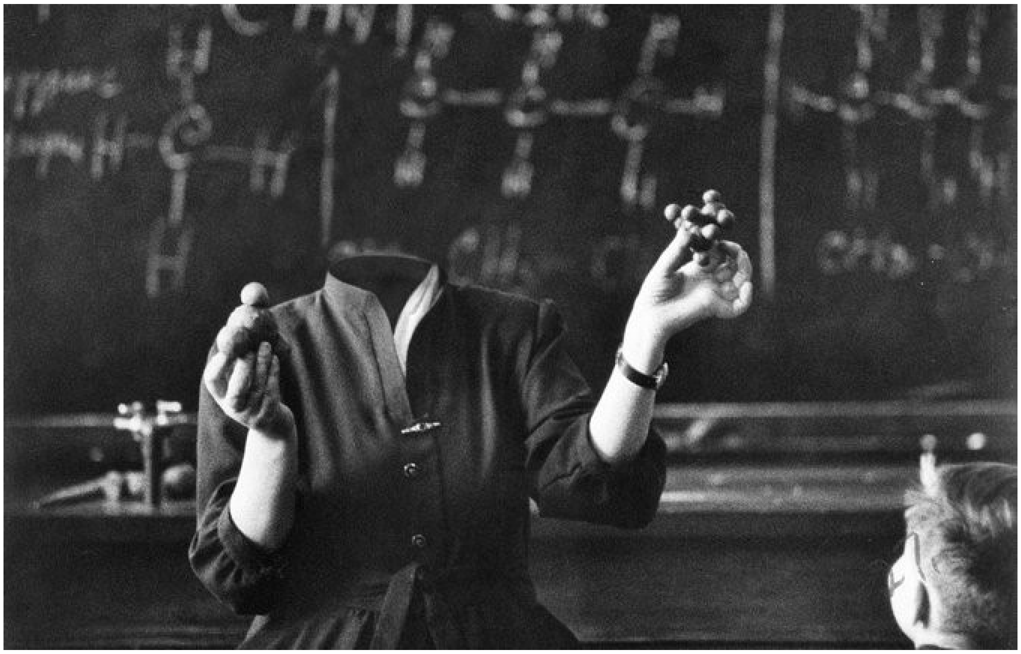

Photo illustration by Derek Brahney. Source photo: Howard Sochurek/The Life Picture Collection/Getty Images.

It was Kosinski's own research, but when he tried to answer that question, he was reduced to a painstaking hunt for clues. At first, he tried covering up or exaggerating parts of faces, trying to see how those changes would affect the machine's predictions. Results were inconclusive. But Kosinski knew that women, in general, have bigger foreheads, thinner jaws and longer noses than men. So he had the computer spit out the 100 faces it deemed most likely to be gay or straight and averaged the proportions of each. It turned out that the faces of gay men exhibited slightly more "feminine" proportions, on average, and that the converse was true for women. If this was accurate, it could support the idea that testosterone levels — already known to mold facial features — help mold sexuality as well.

But it was impossible to say for sure. Other evidence seemed to suggest that the algorithms might also be picking up on culturally driven traits, like straight men wearing baseball hats more often. Or — crucially — they could have been picking up on elements of the photos that humans don't even recognize. "Humans might have trouble detecting these tiny footprints that border on the infinitesimal," Kosinski says. "Computers can do that very easily."

It has become commonplace to hear that machines, armed with machine learning, can outperform humans at decidedly human tasks, from playing Go to playing "Jeopardy!" We assume that is because computers simply have more data-crunching power than our soggy three-pound brains. Kosinski's results suggested something stranger: that artificial intelligences often excel by developing whole new ways of seeing, or even thinking, that are inscrutable to us. It's a more profound version of what's often called the "black box" problem — the inability to discern exactly what machines are doing when they're teaching themselves novel skills — and it has become a central concern in artificial-intelligence research. In many arenas, A.I. methods have advanced with startling speed; deep neural networks can now detect certain kinds of cancer as accurately as a human. But human doctors still have to make the decisions — and they won't trust an A.I. unless it can explain itself.

This isn't merely a theoretical concern. In 2018, the European Union will begin enforcing a law requiring that any decision made by a machine be readily explainable, on penalty of fines that could cost companies like Google and Facebook billions of dollars. The law was written to be powerful and broad and fails to define what constitutes a satisfying explanation or how exactly those explanations are to be reached. It represents a rare case in which a law has managed to leap into a future that academics and tech companies are just beginning to devote concentrated effort to understanding. As researchers at Oxford dryly noted, the law “could require a complete overhaul of standard and widely used algorithmic techniques” — techniques already permeating our everyday lives.

Those techniques can seem inescapably alien to our own ways of thinking. Instead of certainty and cause, A.I. works off probability and correlation. And yet A.I. must nonetheless conform to the society we’ve built — one in which decisions require explanations, whether in a court of law, in the way a business is run or in the advice our doctors give us. The disconnect between how we make decisions and how machines make them, and the fact that machines are making more and more decisions for us, has birthed a new push for transparency and a field of research called explainable A.I., or X.A.I. Its goal is to make machines able to account for the things they learn, in ways that we can understand. But that goal, of course, raises the fundamental question of whether the world a machine sees can be made to match our own.

“Artificial intelligence” is a misnomer, an airy and evocative term that can be shaded with whatever notions we might have about what “intelligence” is in the first place. Researchers today prefer the term “machine learning,” which better describes what makes such algorithms powerful. Let’s say that a computer program is deciding whether to give you a loan. It might start by comparing the loan amount with your income; then it might look at your credit history, marital status or age; then it might consider any number of other data points. After exhausting this “decision tree” of possible variables, the computer will spit out a decision. If the program were built with only a few examples to reason from, it probably wouldn’t be very accurate. But given millions of cases to consider, along with their various outcomes, a machine-learning algorithm could tweak itself — figuring out when to, say, give more weight to age and less to income — until it is able to handle a range of novel situations and reliably predict how likely each loan is to default.

Machine learning isn’t just one technique. It encompasses entire families of them, from “boosted decision trees,” which allow an algorithm to change the weighting it gives to each data point, to “random forests,” which average together many thousands of randomly generated decision trees. The sheer proliferation of different techniques, none of them obviously better than the others, can leave researchers flummoxed over which one to choose. Many of the most powerful are bafflingly opaque; others evade understanding because they involve an avalanche of statistical probability. It can be almost impossible to peek inside the box and see what, exactly, is happening.

Rich Caruana, an academic who works at Microsoft Research, has spent almost his entire career in the shadow of this problem. When he was earning his Ph.D at Carnegie Mellon University in the 1990s, his thesis adviser asked him and a group of others to train a neural net — a forerunner of the deep neural net — to help evaluate risks for patients with pneumonia. Between 10 and 11 percent of cases would be fatal; others would be less urgent, with some percentage of patients recovering just fine without a great deal of medical attention. The problem was figuring out which cases were which — a high-stakes question in, say, an emergency room, where doctors have to make quick decisions about what kind of care to offer. Of all the machine-learning techniques students applied to this question, Caruana’s neural net was the most effective. But when someone on the staff of the University of Pittsburgh Medical Center asked him if they should start using his algorithm, “I said no,” Caruana recalls. “I said we don’t understand what it does inside. I said I was afraid.”

The problem was in the algorithm’s design. Classical neural nets focus only on whether the prediction they gave is right or wrong, tweaking and weighing and recombining all available morsels of data into a tangled web of inferences that seems to get the job done. But some of these inferences could be terrifically wrong. Caruana was particularly concerned by something another graduate student noticed about the data they were handling: It seemed to show that asthmatics with pneumonia fared better than the typical patient. This correlation was real, but the data masked its true cause. Asthmatic patients who contract pneumonia are immediately flagged as dangerous cases; if they tended to fare better, it was because they got the best care the hospital could offer. A dumb algorithm, looking at this data, would have simply assumed asthma meant a patient was likely to get better — and thus concluded that they were in less need of urgent care.

“I knew I could probably fix the program for asthmatics,” Caruana says. “But what else did the neural net learn that was equally wrong? It couldn’t warn me about the unknown unknowns. That tension has bothered me since the 1990s.”

The story of asthmatics with pneumonia eventually became a legendary allegory in the machine-learning community. Today, Caruana is one of perhaps a few dozen researchers in the United States dedicated to finding more transparent new approaches to machine learning. For the last six years, he has been creating a new model that combines a number of machine-learning techniques. The result is as accurate as his original neural network, and it can spit out charts that show how each individual variable — from asthma to age — is predictive of mortality risk, making it easier to see which ones exhibit particularly unusual behavior. Immediately, asthmatics are revealed as a far outlier. Other strange truths surface, too: For example, risk for people age 100 goes down suddenly. “If you made it to this round number of 100,” Caruana says, “it seemed as if the doctors were saying, ‘Let’s try to get you another year,’ which might not happen if you’re 93.”

Caruana may have brought clarity to his own project, but his solution only underscored the fact the explainability is a kaleidoscopic problem. The explanation a doctor needs from a machine isn’t the same as the one a fighter pilot might need or the one an N.S.A. analyst sniffing out a financial fraud might need. Different details will matter, and different technical means will be needed for finding them. You couldn’t, for example, simply use Caruana’s techniques on facial data, because they don’t apply to image recognition. There may, in other words, eventually have to be as many approaches to explainability as there are approaches to machine learning itself.

Three years ago, David Gunning, one of the most consequential people in the emerging discipline of X.A.I., attended a brainstorming session at a state university in North Carolina. The event had the title “Human-Centered Big Data,” and it was sponsored by a government-funded think tank called the Laboratory for Analytic Sciences. The idea was to connect leading A.I. researchers with experts in data visualization and human-computer interaction to see what new tools they might invent to find patterns in huge sets of data. There to judge the ideas, and act as hypothetical users, were analysts for the C.I.A., the N.S.A. and sundry other American intelligence agencies.

The researchers in Gunning's group stepped confidently up to the white board, showing off new, more powerful ways to draw predictions from a machine and then visualize them. But the intelligence analyst evaluating their pitches, a woman who couldn't tell anyone in the room what she did or what tools she was using, waved it all away. Gunning remembers her as plainly dressed, middle-aged, typical of the countless government agents he had known who toiled thanklessly in critical jobs. "None of this solves my problem," she said. "I don't need to be able to visualize another recommendation. If I'm going to sign off on a decision, I need to be able to justify it." She was issuing what amounted to a broadside. It wasn't just that a clever graph indicating the best choice wasn't the same as explaining why that choice was correct. The analyst was pointing to a legal and ethical motivation for explainability: Even if a machine made perfect decisions, a human would still have to take responsibility for them — and if the machine's rationale was beyond reckoning, that could never happen.

Gunning, a grandfatherly military man whose buzz cut has survived his stints as a civilian, is a program manager at the Defense Advanced Research Projects Agency. He works in Darpa's shiny new midrise tower in downtown Alexandria, Va. — an office indistinguishable from the others nearby, except that the security guard out front will take away your cellphone and warn you that turning on the Wi-Fi on your laptop will make security personnel materialize within 30 seconds. Darpa managers like Gunning don't have permanent jobs; the expectation is that they serve four-year "tours," dedicated to funding cutting-edge research along a single line of inquiry. When he found himself at the brainstorming session, Gunning had recently completed his second tour as a sort of Johnny Appleseed for A.I.: Starting in the 1990s, he has founded hundreds of projects, from the first application of machine-learning techniques to the internet, which presaged the first search engines, to the project that eventually spun off as Siri, Apple's voicecontrolled assistant. “I’m proud to be a dinosaur,” he says with a smile.

As of now, most of the military’s practical applications of such technology involve performing enormous calculations beyond the reach of human patience, like predicting how to route supplies. But there are more ambitious applications on the horizon. One recent research program tried to use machine learning to sift through millions of video clips and internet messages in Yemen to detect cease-fire violations; if the machine does find something, it has to be able to describe what’s worth paying attention to. Another pressing need is for drones flying on self-directed missions to be able to explain their limitations so that the humans commanding the drones know what the machines can — and cannot — be asked to do. Explainability has thus become a hurdle for a wealth of possible projects, and the Department of Defense has begun to turn its eye to the problem.

After that brainstorming session, Gunning took the analyst’s story back to Darpa and soon signed up for his third tour. As he flew across the country meeting with computer scientists to help design an overall strategy for tackling the problem of X.A.I., what became clear was that the field needed to collaborate more broadly and tackle grander problems. Computer science, having leapt beyond the bounds of considering purely technical problems, had to look further afield — to experts, like cognitive scientists, who study the ways humans and machines interact.

This represents a full circle for Gunning, who began his career as a cognitive psychologist working on how to design better automated systems for fighter pilots. Later, he began working on what’s now called “old-fashioned A.I.” — so-called expert systems in which machines were given voluminous lists of rules, then tasked with drawing conclusions by recombining those rules. None of those efforts was particularly successful, because it was impossible to give the computer a set of rules long enough, or flexible enough, to approximate the power of human reasoning. A.I.’s current blossoming came only when researchers began inventing new techniques for letting machines find their own patterns in the data.

Gunning’s X.A.I. initiative, which kicked off this year, provides $75 million in funding to 12 new research programs; by the power of the purse strings, Gunning has refocused the energies of a significant part of the American A.I. research community. His hope is that by making these new A.I. methods accountable to the demands of human psychology, they will become both more useful and more powerful. “The real secret is finding a way to put labels on the concepts inside a deep neural net,” he says. If the concepts inside can be labeled, then they can be used for reasoning — just like those expert systems were supposed to do in A.I.’s first wave.

Deep neural nets, which evolved from the kinds of techniques that Rich Caruana was experimenting with in the 1990s, are now the class of machine learning that seems most opaque. Just like old-fashioned neural nets, deep neural networks seek to draw a link between an input on one end (say, a picture from the internet) and an output on the other end (“This is a picture of a dog”). And just like those older neural nets, they consume all the examples you might give them, forming their own webs of inference that can then be applied to pictures they’ve never seen before. Deep neural nets remain a hotbed of research because they have produced some of the most breathtaking technological accomplishments of the last decade, from learning how to translate words with better-than-human accuracy to learning how to drive.

To create a neural net that can reveal its inner workings, the researchers in Gunning’s portfolio are pursuing a number of different paths. Some of these are technically ingenious — for example, designing new kinds of deep neural networks made up of smaller, more easily understood modules, which can fit together like Legos to accomplish complex tasks. Others involve psychological insight: One team at Rutgers is designing a deep neural network that, once it makes a decision, can then sift through its data set to find the example that best demonstrates why it made that decision. (The idea is partly inspired by psychological studies of real-life experts like firefighters, who don’t clock in for a shift thinking, These are the 12 rules for fighting fires; when they see a fire before them, they compare it with ones they’ve seen before and act accordingly.) Perhaps the most ambitious of the dozen different projects are those that seek to bolt new explanatory capabilities onto existing deep neural networks. Imagine giving your pet dog the power of speech, so that it might finally explain what’s so interesting about squirrels. Or, as Trevor Darrell, a lead investigator on one of those teams, sums it up, “The solution to explainable A.I. is more A.I.”

Five years ago, Darrell and some colleagues had a novel idea for letting an A.I. teach itself how to describe the contents of a picture. First, they created two deep neural networks: one dedicated to image recognition and another to translating languages. Then they lashed these two together and fed them thousands of images that had captions attached to them. As the first network learned to recognize the objects in a picture, the second simply watched what was happening in the first, then learned to associate certain words with the activity it saw. Working together, the two networks could identify the features of each picture, then label them. Soon after, Darrell was presenting some different work to a group of computer scientists when someone in the audience raised a hand, complaining that the techniques he was describing would never be explainable. Darrell, without a second thought, said, Sure — but you could make it explainable by once again lashing two deep neural networks together, one to do the task and one to describe it.

Darrell’s previous work had piggybacked on pictures that were already captioned. What he was now proposing was creating a new data set and using it in a novel way. Let’s say you had thousands of videos of baseball highlights. An image-recognition network could be trained to spot the players, the ball and everything happening on the field, but it wouldn’t have the words to label what they were. But you might then create a new data set, in which volunteers had written sentences describing the contents of every video. Once combined, the two networks should then be able to answer queries like “Show me all the double plays involving the Boston Red Sox” — and could potentially show you what cues, like the logos on uniforms, it used to figure out who the Boston Red Sox are.

Call it the Hamlet strategy: lending a deep neural network the power of internal monologue, so that it can narrate what’s going on inside. But do the concepts that a network has taught itself align with the reality that humans are describing, when, for example, narrating a baseball highlight? Is the network recognizing the Boston Red Sox by their logo or by some other obscure signal, like “median facial-hair distribution,” that just happens to correlate with the Red Sox? Does it actually have the concept of “Boston Red Sox” or just some other strange thing that only the computer understands? It's an ontological question: Is the deep neural network really seeing a world that corresponds to our own?

We human beings seem to be obsessed with black boxes: The highest compliment we give to technology is that it feels like magic. When the workings of a new technology is too obvious, too easy to explain, it can feel banal and uninteresting. But when I asked David Jensen — a professor at the University of Massachusetts at Amherst and one of the researchers being funded by Gunning — why X.A.I. had suddenly become a compelling topic for research, he sounded almost soulful: "We want people to make informed decisions about whether to trust autonomous systems," he said. "If you don't, you're depriving people of the ability to be fully independent human beings."

A decade in the making, the European Union's General Data Protection Regulation finally goes into effect in May 2018. It's a sprawling, many-tentacled piece of legislation whose opening lines declare that the protection of personal data is a universal human right. Among its hundreds of provisions, two seem aimed squarely at where machine learning has already been deployed and how it's likely to evolve. Google and Facebook are most directly threatened by Article 21, which affords anyone the right to opt out of personally tailored ads. The next article then confronts machine learning head on, limning a so-called right to explanation: E.U. citizens can contest "legal or similarly significant" decisions made by algorithms and appeal for human intervention. Taken together, Articles 21 and 22 introduce the principle that people are owed agency and understanding when they're faced by machine-made decisions.

For many, this law seems frustratingly vague. Some legal scholars argue that it might be toothless in practice. Others claim that it will require the basic workings of Facebook and Google to change, lest they face penalties of 4 percent of their revenue. It remains to be seen whether complying with the law will mean a heap of fine print and an extra check box buried in a pop-up window, some new kind of warning-label system marking every machine-made decision or much more profound changes.

If Google is one of the companies most endangered by this new scrutiny on A.I., it's also the company with the greatest wherewithal to lead the whole industry in solving the problem. Even among the company's astonishing roster of A.I. talent, one particular star is Chris Olah, who holds the title of research scientist — a title shared by Google's many ex-professors and Ph.D.s — without ever having completed more than a year of college. Olah has been working for the last couple of years on creating new ways to visualize the inner workings of a deep neural network. You might recall when Google created a hallucinatory tool called Deep Dream, which produced psychedelic distortions when you fed it an image and which went viral when people used it to create hallucinatory mash-ups like a doll covered in a pattern of doll eyes and a portrait of Vincent Van Gogh made up in places of bird beaks. Olah was one of many Google researchers on the team, led by Alex Mordvintsev, that worked on Deep Dream. It may have seemed like a folly, but it was actually a technical steppingstone.

Olah speaks faster and faster as he sinks into an idea, and the words tumbled out of him almost too quickly to follow as he explained what he found so exciting about the work he was doing. "The truth is, it's really beautiful. There's some sense in which we don't know what it means to see. We don't understand how humans do it," he told me, hands gesturing furiously. "We want to understand something not just about neural nets but something deeper about reality." Olah's hope is that deep neural networks reflect something deeper about parsing data — that insights gleaned from them might in turn shed light on how our brains work.

Olah showed me a sample of work he was preparing to publish with a set of collaborators, including Mordvintsev; it was made public this month. The tool they had developed was basically an ingenious way of testing a deep neural network. First, it fed the network a random image of visual noise. Then it tweaked that image over and over again, working to figure out what excited each layer in the network the most. Eventually, that process would find the platonic ideal that each layer of the network was searching for. Olah demonstrated with a network trained to classify different breeds of dogs. You could pick out a neuron from the topmost layer while it was analyzing a picture of a golden retriever. You could see the ideal it was looking for — in this case, a hallucinatory mash-up of floppy ears and a forlorn expression. The network was indeed homing in on higher-level traits that we could understand.

Watching him use the tool, I realized that it was exactly what the psychologist Michal Kosinski needed — a key to unlock what his deep neural network was seeing when it categorized profile pictures as gay or straight. Kosinski's most optimistic view of his research was that it represented a new kind of science in which machines could access truths that lay beyond human intuition. The problem was reducing what a computer knew into a single conclusion that a human could grasp and consider. He had painstakingly tested his data set by hand and found evidence that the computer might be discovering hormonal signals in facial structure. That evidence was still fragmentary. But with the tool that Olah showed me, or one like it, Kosinski might have been able to pull back the curtain on how his mysterious A.I. was working. It would be as obvious and intuitive as a picture the computer had drawn on its own.