HOW TO TEACH ARTIFICIAL INTELLIGENCE SOME COMMON SENSE
SHARE
1396
FIVE YEARS AGO, the coders at DeepMind, a London-based artificial intelligence company, watched excitedly as an AI taught itself to play a classic arcade game. They'd used the hot technique of the day, deep learning, on a seemingly whimsical task: mastering Breakout,1 the Atari game in which you bounce a ball at a wall of bricks, trying to make each one vanish.

1 Steve Jobs was working at Atari when he was commissioned to create 1976's Breakout, a job no other engineer wanted. He roped his friend Steve Wozniak, then at Hewlett-Packard, into helping him.

Deep learning is self-education for machines; you feed an AI huge amounts of data, and eventually it begins to discern patterns all by itself. In this case, the data was the activity on the screen—blocky pixels representing the bricks, the ball, and the player's paddle. The DeepMind AI, a so-called neural network made up of layered algorithms, wasn't programmed with any knowledge about how Breakout works, its rules, its goals, or even how to play it. The coders just let the neural net examine the results of each action, each bounce of the ball. Where would it lead?

To some very impressive skills, it turns out. During the first few games, the AI flailed around. But after playing a few hundred times, it had begun accurately bouncing the ball. By the 600th game, the neural net was using a more expert move employed by human Breakout players, chipping through an entire column of bricks and setting the ball bouncing merrily along the top of the wall.

"That was a big surprise for us," Demis Hassabis, CEO of DeepMind, said at the time. "The strategy completely emerged from the underlying system." The AI had shown itself capable of what seemed to be an unusually subtle piece of humanlike thinking, a grasping of the inherent concepts behind Breakout. Because neural nets loosely mirror the structure of the human brain, the theory was that they should mimic, in some respects, our own style of cognition. This moment seemed to serve as proof that the theory was right.

December 2018. Subscribe to WIRED. AXIS OF STRENGTH
Then, last year, computer scientists at Vicarious, an AI firm in San Francisco, offered an interesting reality check. They took an AI like the one used by DeepMind and trained it on Breakout. It played great. But then they slightly tweaked the layout of the game. They lifted the paddle up higher in one iteration; in another, they added an unbreakable area in the center of the blocks.

A human player would be able to quickly adapt to these changes; the neural net couldn't. The seemingly supersmart AI could play only the exact style of Breakout it had spent hundreds of games mastering. It couldn't handle something new.

"We humans are not just pattern recognizers," Dileep George, a computer scientist who cofounded Vicarious, tells me. "We're also building models about the things we see. And these are causal models—we understand about cause and effect." Humans engage in reasoning, making logical inferences about the world around us; we have a store of common-sense knowledge that helps us figure out new situations. When we see a game of Breakout that's a little different from the one we just played, we realize it's likely to have mostly the same rules and goals. The neural net, on the other hand, hadn't understood anything about Breakout. All it could do was follow the pattern. When the pattern changed, it was helpless.

THE A.I. ISSUE

TOM SIMONITE
The DIY Tinkerers Harnessing the Power of Artificial Intelligence

JESSI HEMPEL
Fei-Fei Li's Quest to Make AI Better for Humanity

SHAUN RAVIV
The Genius Neuroscientist Who Might Hold the Key to True AI

Deep learning is the reigning monarch of AI. In the six years since it exploded into the mainstream, it has become the dominant way to help machines sense and perceive the world around them. It powers Alexa's speech recognition, Waymo's self-driving cars, and Google's on-the-fly translations. Uber is in some respects a giant optimization problem, using machine learning to figure out where riders will need cars. Baidu, the Chinese tech giant, has more than 2,000 engineers cranking away on neural net AI. For years, it seemed as though deep learning would only keep getting better, leading inexorably to a machine with the fluid, supple intelligence of a person.

But some heretics argue that deep learning is hitting a wall. They say that, on its own, it'll never produce generalized intelligence, because truly humanlike intelligence isn't just pattern recognition. We need to start figuring out how to imbue AI with everyday common sense, the stuff of human smarts. If we don't, they warn, we'll keep bumping up against the limits of deep learning, like visual-recognition systems that can be easily fooled by changing a few inputs, making a deep-learning model think a turtle is a gun. But if we succeed, they say, we'll witness an explosion of safer, more useful devices—health care robots that navigate a cluttered home, fraud detection systems that don't trip on false positives, medical breakthroughs powered by machines that ponder cause and effect in disease.

But what does true reasoning look like in a machine? And if deep learning can't get us there, what can?

BETH HOLZER
GARY MARCUS IS a pensive, bespectacled 48-year-old professor of psychology and neuroscience at New York University, and he's probably the most famous apostate of orthodox deep learning.

Marcus first got interested in artificial intelligence in the 1980s and '90s, when neural nets were still in their experimental phase, and he's been making the same argument ever since. "It's not like I came to this party late and want to pee on it," Marcus told me when I met him at his apartment

near NYU. (We are also personal friends.) “As soon as deep learning erupted, I said ‘This is the wrong direction, guys!’ ”

Back then, the strategy behind deep learning was the same as it is today. Say you wanted a machine to teach itself to recognize daisies. First you’d code some algorithmic “neurons,” connecting them in layers like a sandwich (when you use several layers, the sandwich gets thicker or deep—hence “deep” learning). You’d show an image of a daisy to the first layer, and its neurons would fire or not fire based on whether the image resembled the examples of daisies it had seen before. The signal would move on to the next layer, where the process would be repeated. Eventually, the layers would winnow down to one final verdict.

At first, the neural net is just guessing blindly; it starts life a blank slate, more or less. The key is to establish a useful feedback loop. Every time the AI misses a daisy, that set of neural connections weakens the links that led to an incorrect guess; if it’s successful, it strengthens them. Given enough time and enough daisies, the neural net gets more accurate. It learns to intuit some pattern of daisy-ness that lets it detect the daisy (and not the sunflower or aster) each time. As the years went on, this core idea—start with a naive network and train by repetition—was improved upon and seemed useful nearly anywhere it was applied.

But Marcus was never convinced. For him, the problem is the blank slate: It assumes that humans build their intelligence purely by observing the world around them, and that machines can too. But Marcus doesn’t think that’s how humans work. He walks the intellectual path laid down by Noam Chomsky,2 who argued that humans are born wired to learn, programmed to master language and interpret the physical world.

2 In 1975 the psychologist Jean Piaget and the linguist Noam Chomsky met in France for what would prove to be a historic debate. Grossly simplified, Piaget argued that human brains are blank-slate self-learning

machines, and Chomsky that they are endowed with some preprogrammed smarts.

For all their supposed braininess, he notes, neural nets don't appear to work the way human brains do. For starters, they're much too data-hungry. In most cases, each neural net requires thousands or millions of examples to learn from. Worse, each time you want a neural net to recognize a new type of item, you have to start from scratch. A neural net trained to recognize only canaries isn't of any use in recognizing, say, birdsong or human speech.

"We don't need massive amounts of data to learn," Marcus says. His kids didn't need to see a million cars before they could recognize one. Better yet, they can generalize; when they see a tractor for the first time, they understand that it's sort of like a car. They can also engage in counterfactuals. Google Translate can map the French equivalent of the English sentence "The glass was pushed, so it fell off the table." But it doesn't know what the words mean, so it couldn't tell you what would happen if the glass weren't pushed. Humans, Marcus notes, grasp not just the patterns of grammar but the logic behind it. You could give a young child a fake verb like pilk, and she'd likely be able to reason that the past tense would be pilked. She hasn't seen that word before, of course. She hasn't been "trained" on it. She has just intuited some logic about how language works and can apply it to a new situation.

"These deep-learning systems don't know how to integrate abstract knowledge," says Marcus, who founded a company that created AI to learn with less data (and sold the company to Uber in 2016).

Earlier this year, Marcus published a white paper on arXiv, arguing that, without some new approaches, deep learning might never get past its current limitations. What it needs is a boost—rules that supplement or are built in to help it reason about the world.

BETH HOLZER

OREN ETZIONI IS a smiling bear of a guy. A computer scientist who runs the Allen Institute for Artificial Intelligence in Seattle, he greets me in his

bright office wearing jeans and a salmon-colored shirt, ushering me in past a whiteboard scrawled with musings about machine intelligence. (“DEFINE SUCCESS,” “WHAT’S THE TASK?”) Outside, in the sun-drenched main room of the institute, young AI researchers pad around sylphlike, headphones attached, quietly pecking at keyboards.

Etzioni and his team are working on the common-sense problem. He defines it in the context of two legendary AI moments—the trouncing of the chess grandmaster Garry Kasparov3 by IBM’s Deep Blue in 1997 and the equally shocking defeat of the world’s top Go player by DeepMind’s AlphaGo last year. (Google bought DeepMind in 2014.)

3 In 1996, Kasparov—then the best chess player in the world—beat Deep Blue. During a rematch a year later, Kasparov surrendered after 19 moves. He later told a reporter: “I’m a human being. When I see something that is well beyond my understanding, I’m afraid.”

“With Deep Blue we had a program that would make a superhuman chess move—while the room was on fire,” Etzioni jokes. “Right? Completely lacking context. Fast-forward 20 years, we’ve got a computer that can make a superhuman Go move—while the room is on fire.” Humans, of course, do not have this limitation. His team plays weekly games of bughouse chess, and if a fire broke out the humans would pull the alarm and run for the doors.

Humans, in other words, possess a base of knowledge about the world (fire burns things) mixed with the ability to reason about it (you should try to move away from an out-of-control fire). For AI to truly think like people, we need to teach it the stuff that everyone knows, like physics (balls tossed in the air will fall) or the relative sizes of things (an elephant can’t fit in a bathtub). Until AI possesses these basic concepts, Etzioni figures, it won’t be able to reason.

With an infusion of hundreds of millions of dollars from Paul Allen,4 Etzioni and his team are trying to develop a layer of common-sense reasoning to work with the existing style of neural net. (The Allen Institute is a nonprofit, so everything they discover will be published, for anyone to use.)

4 Microsoft cofounder and philanthropist Paul Allen donated billions to science, climate, and health research, as well as to Seattle causes. He died of complications from cancer on October 15 at age 65.

The first problem they face is answering the question, What is common sense?

Etzioni describes it as all the knowledge about the world that we take for granted but rarely state out loud. He and his colleagues have created a set of benchmark questions that a truly reasoning AI ought to be able to answer: If I put my socks in a drawer, will they be there tomorrow? If I stomp on someone's toe, will they be mad?

One way to get this knowledge is to extract it from people. Etzioni's lab is paying crowdsourced humans on Amazon Mechanical Turk to help craft common-sense statements. The team then uses various machine-learning techniques—some old-school statistical analyses, some deep-learning neural nets—to draw lessons from those statements. If they do it right, Etzioni believes they can produce reusable Lego bricks of computer reasoning: One set that understands written words, one that grasps physics, and so on.

Yejin Choi, one of Etzioni's leading common-sense scientists, has led several of these crowdsourced efforts. In one project, she wanted to develop an AI that would understand the intent or emotion implied by a person's actions or statements. She started by examining thousands of online stories, blogs, and idiom entries in Wiktionary and extracting "phrasal events," such as the sentence "Jeff punches Roger's lights out." Then she'd anonymize each phrase—"Person X punches Person Y's lights out"—and ask the Turkers to describe the intent of Person X: Why did they do that? When she had gathered 25,000 of these marked-up sentences, she used them to train a machine-learning system to analyze sentences it had never seen before and infer the emotion or intent of the subject.

LEARN MORE

THE WIRED GUIDE TO ARTIFICIAL INTELLIGENCE

At best, the new system worked only half the time. But when it did, it evinced some very humanlike perception: Given a sentence like “Oren cooked Thanksgiving dinner,” it predicted that Oren was trying to impress his family. “We can also reason about others’ reactions, even if they’re not mentioned,” Choi notes. “So X’s family probably feel impressed and loved.” Another system her team built used Turkers to mark up the psychological states of people in stories; the resulting system could also draw some sharp inferences when given a new situation. It was told, for instance, about a music instructor getting angry at his band’s lousy performance and that “the instructor was furious and threw his chair.” The AI predicted that the musicians would “feel fear afterwards,” even though the story doesn’t explicitly say so.

Choi, Etzioni, and their colleagues aren’t abandoning deep learning. Indeed, they regard it as a very useful tool. But they don’t think there is a shortcut to the laborious task of coaxing people to explicitly state the weird, invisible, implied knowledge we all possess. Deep learning is garbage in, garbage out. Merely feeding a neural net tons of news articles isn’t enough, because it wouldn’t pick up on the unstated knowledge, the obvious stuff that writers didn’t bother to mention. As Choi puts it, “People don’t say ‘My house is bigger than me.’ ” To help tackle this problem, she had the Turkers analyze the physical relationships implied by 1,100 common verbs, such as “X threw Y.” That, in turn, allowed for a simple statistical model that could take the sentence “Oren threw the ball” and infer that the ball must be smaller than Oren.

Another challenge is visual reasoning. Aniruddha Kembhavi, another of Etzioni’s AI scientists, shows me a virtual robot wandering around an onscreen house. Other Allen Institute scientists built the Sims-like house, filling it with everyday items and realistic physics—kitchen cupboards full of dishes, couches that can be pushed around. Then they designed the robot, which looks like a dark gray garbage canister with arms, and told it to hunt down certain items. After thousands of tasks, the neural net gains a basic grounding in real-life facts.

“What this agent has learned is, when you ask it ‘Do I have tomatoes?’ it doesn’t go and open all the cabinets. It prefers to open the fridge,” Kembhavi says. “Or if you say ‘Find me my keys,’ it doesn’t try to pick up

the television. It just looks behind the television. It has learned that TVs aren't usually picked up."

Etzioni and his colleagues hope that these various components—Choi's language reasoning, the visual thinking, other work they're doing on getting an AI to grasp textbook science information—can all eventually be combined. But how long will it take, and what will the final products look like? They don't know. The common-sense systems they're building still make mistakes, sometimes more than half the time. Choi estimates she'll need around a million crowdsourced human statements as she trains her various language-parsing AIs. Building common sense, it would seem, is uncommonly hard.

THERE ARE OTHER pathways to making machines that reason, and they're even more labor-intensive. For example, you could simply sit down and write out, by hand, all the rules that tell a machine how the world works. This is how Doug Lenat's Cyc project works. For 34 years, Lenat has employed a team of engineers and philosophers to code 25 million rules of general common sense, like "water is wet" or "most people know the first names of their friends." This lets Cyc deduce things: "Your shirt is wet, so you were probably in the rain." The advantage is that Lenat has exquisite control over what goes into Cyc's database; that isn't true of crowdsourced knowledge.

Brute-force, handcrafted AI has become unfashionable in the world of deep learning. That's partly because it can be "brittle": Without the right rules about the world, the AI can get flummoxed. This is why scripted chatbots are so frustrating; if they haven't been explicitly told how to answer a question, they have no way to reason it out. Cyc is enormously more capable than a chatbot and has been licensed for use in health care systems, financial services, and military projects. But the work is achingly slow, and it's expensive. Lenat says it has cost around $200 million to develop Cyc.

But a bit of hand coding could be how you replicate some of the built-in knowledge that, according to the Chomskyite view, human brains possess. That's what Dileep George and the Vicarious researchers did with Breakout. To create an AI that wouldn't get stumped by changes to the

layout of the game, they abandoned deep learning and built a system that included hard-coded basic assumptions. Without too much trouble, George tells me, their AI learned “that there are objects, and there are interactions between objects, and that the motion of one object can be causally explained between the object and something else.”

As it played Breakout, the system developed the ability to weigh different courses of action and their likely outcomes. This worked in reverse too. If the AI wanted to break a block in the far left corner of the screen, it reasoned to put the paddle in the far right corner. Crucially, this meant that when Vicarious changed the layout of the game—adding new bricks or raising the paddle—the system compensated. It appeared to have extracted some general understanding about Breakout itself.

Granted, there are trade-offs in this type of AI engineering. It’s arguably more painstaking to craft and takes careful planning to figure out precisely what foreordained logic to feed into the system. It’s also hard to strike the right balance of speed and accuracy when designing a new system. George says he looks for the minimum set of data “to put into the model so it can learn quickly.” The fewer assumptions you need, the more efficiently the machine will make decisions. Once you’ve trained a deep-learning model to recognize cats, you can show it a Russian blue it has never seen and it renders the verdict—it’s a cat!—almost instantaneously. Having processed millions of photos, it knows not only what makes a cat a cat but also the fastest way to identify one. In contrast, Vicarious’ style of AI is slower, because it’s actively making logical inferences as it goes.

When the Vicarious AI works well, it can learn from much less data. George’s team created an AI to bust captchas,5 those “I’m not a robot” obstacles online, by recognizing characters in spite of their distorted, warped appearance.

5 Captcha stands for “Completely Automated Public Turing test to tell Computers and Humans Apart.” It originated at Carnegie Mellon University in 2000; Yahoo was the first big company to make its use commonplace.

Much as with the Breakout system, they endowed their AI with some abilities up front, such as knowledge that helps it discern the likely edges of

characters. With that bootstrapping in place, they only needed to train the AI on 260 images before it learned to break captchas with 90.4 percent accuracy. In contrast, a neural net needed to be trained on more than 2.3 million images before it could break a captcha.

Others are building common-sense-like structure into neural nets in different ways. Two researchers at DeepMind, for instance, recently created a hybrid system—part deep learning, part more traditional techniques—known as inductive logic programming. The goal was to produce something that could do mathematical reasoning.

They trained it on the children's game fizz-buzz, in which you count upward from 1, saying "fizz" if a number is divisible by 3 and "buzz" if it is divisible by 5. A regular neural net would be able to do this only for numbers it had seen before; train it up to 100 and it would know that 99 is "fizz" and 100 is "buzz." But it wouldn't know what to do with 105. In contrast, the hybrid DeepMind system seemed to understand the rule and went past 100 with no problem. Edward Grefenstette, one of the DeepMind coders who built the hybrid, says, "You can train systems that will generalize in a way that deep-learning networks simply couldn't on their own."

BETH HOLZER

Yann LeCun, a deep-learning pioneer and the current head of Facebook's AI research wing, agrees with many of the new critiques of the field. He acknowledges that it requires too much training data, that it can't reason, that it doesn't have common sense. "I've been basically saying this over and over again for the past four years," he reminds me. But he remains steadfast that deep learning, properly crafted, can provide the answer. He disagrees with the Chomskyite vision of human intelligence. He thinks human brains develop the ability to reason solely through interaction, not built-in rules. "If you think about how animals and babies learn, there's a lot of things that are learned in the first few minutes, hours, days of life that seem to be done so fast that it looks like they are hardwired," he notes. "But in fact they don't need to be hardwired, because they can be learned so quickly." In this view, to figure out the physics of the world, a baby just moves its head around, data-crunches the incoming imagery, and concludes that, hey, depth of field is a thing.

Still, LeCun admits it's not yet clear which routes will help deep learning get past its humps. It might be "adversarial" neural nets, a relatively new technique in which one neural net tries to fool another neural net with fake data—forcing the second one to develop extremely subtle internal representations of pictures, sounds, and other inputs. The advantage here is that you don't have the "data hungriness" problem. You don't need to collect millions of data points on which to train the neural nets, because they're learning by studying each other. (Apocalyptic side note: A similar method is being used to create those profoundly troubling "deepfake" videos in which someone appears to be saying or doing something they are not.)

I met LeCun at the offices of Facebook's AI lab in New York. Mark Zuckerberg recruited him in 2013, with the promise that the lab's goal would be to push the limits of ambitious AI, not just produce minor tweaks for Facebook's products. Like an academic lab, LeCun and his researchers publish their work for others to access.

LeCun, who retains the rich accent of his native France and has a Bride of Frankenstein shock of white in his thick mass of dark hair, stood at a whiteboard energetically sketching out theories of possible deep-learning advances. On the facing wall was a set of gorgeous paintings from Stanley Kubrick's 2001: A Space Odyssey—the main spaceship floating in deep space, the wheel-like ship orbiting Earth. "Oh, yes," LeCun said, when I pointed them out; they were reprints of artwork Kubrick commissioned for the movie.

It was weirdly unsettling to discuss humanlike AI with those images around, because of course HAL 9000,6 the humanlike AI in 2001, turns out to be a highly efficient murderer.

6 HAL was originally supposed to be voiced by Martin Balsam, an actor with a thick Bronx accent. After recording, however, director Stanley Kubrick decided Balsam sounded "too colloquially American." He was replaced by Canadian actor Douglas Rain.

And this pointed to a deeper philosophical question that floats over the whole debate: Is smarter AI even a good idea? Vicarious' system cracked captcha, but the whole point of captcha is to prevent bots from impersonating humans. Some AI thinkers worry that the ability to talk to humans and understand their psychology could make a rogue AI incredibly dangerous. Nick Bostrom7 at the University of Oxford has sounded the alarm about the dangers of creating a "superintelligence," an AI that self-improves and rapidly outstrips humanity, able to outthink and outflank us in every way. (One way he suggests it might amass control is by manipulating people—something for which possessing a "theory of mind" would be quite useful.)

7 In 2003, Bostrom published the now-famous paper-clip warning about superintelligence: "A well-meaning team of programmers [could] make a big mistake in designing its goal system. This could result … in a super-intelligence whose top goal is the manufacturing of paper clips, with the consequence that it starts transforming first all of Earth and then increasing portions of space into paper-clip manufacturing facilities."

Elon Musk is sufficiently convinced of this danger that he has funded OpenAI, an organization dedicated to the notion of safe AI.

This future doesn't keep Etzioni up at night. He's not worried about AI becoming maliciously superintelligent. "We're worried about something taking over the world," he scoffs, "that can't even on its own decide to play chess again." It's not clear how an AI would develop a desire to do so or what that desire would look like in software. Deep learning can conquer chess, but it has no inborn will to play.

What does concern him is that current AI is woefully inept. So while we might not be creating HAL with a self-preserving intelligence, an "inept AI attached to deadly weapons can easily kill," he says. This is partly why Etzioni is so determined to give AI some common sense. Ultimately, he argues, it will make AI safer; the idea that humanity shouldn't be wholesale slaughtered is, of course, arguably a piece of common-sense knowledge itself. (Part of the Allen Institute's mandate is to make AI safer by making it more reasonable.)

RELATED STORIES

TOM SIMONITE
Researchers Call for More Humanity in Artificial Intelligence

LOUISE MATSAKIS
To Break a Hate-Speech Detection Algorithm, Try 'Love'

NICHOLAS THOMPSON
Emmanuel Macron Talks to WIRED About France's AI Strategy

Etzioni notes that the dystopic sci-fi visions of AI are less risky than near-term economic displacement. The better AI gets at common sense, the more rapidly it'll take over jobs that currently are too hard for mere pattern--matching deep learning: drivers, cashiers, managers, analysts of all stripes, and even (alas) journalists. But truly reasoning AI could wreak havoc even beyond the economy. Imagine how good political disinformation bots would be if they could use common-sense knowledge to appear indistinguishably human on Twitter or Facebook or in mass phone calls.

Marcus agrees that reasoning AI will have dangers. But the upsides, he says, would be huge. AI that could reason and perceive like humans yet move at the speed of computers could revolutionize science, teasing out causal connections at a pace impossible for us alone. It could follow if-then chains and ponder counterfactuals, running mental experiments the way humans do, except with massive robotic knowledge. "We might finally be able to cure mental illness, for example," Marcus adds. "AI might be able to understand these complex biological cascades of proteins that are involved in building brains and having them work correctly or not."

Sitting beneath the images from 2001, LeCun makes a bit of a heretical point himself. Sure, making artificial intelligence more humanlike helps AI to navigate our world. But directly replicating human styles of thought? It's not clear that'd be useful. We already have humans who can think like humans; maybe the value of smart machines is that they are quite alien from us.

“They will tend to be more useful if they have capabilities we don’t have,” he tells me. “Then they’ll become an amplifier for intelligence. So to some extent you want them to have a nonhuman form of intelligence ... You want them to be more rational than humans.” In other words, maybe it’s worth keeping artificial intelligence a little bit artificial.