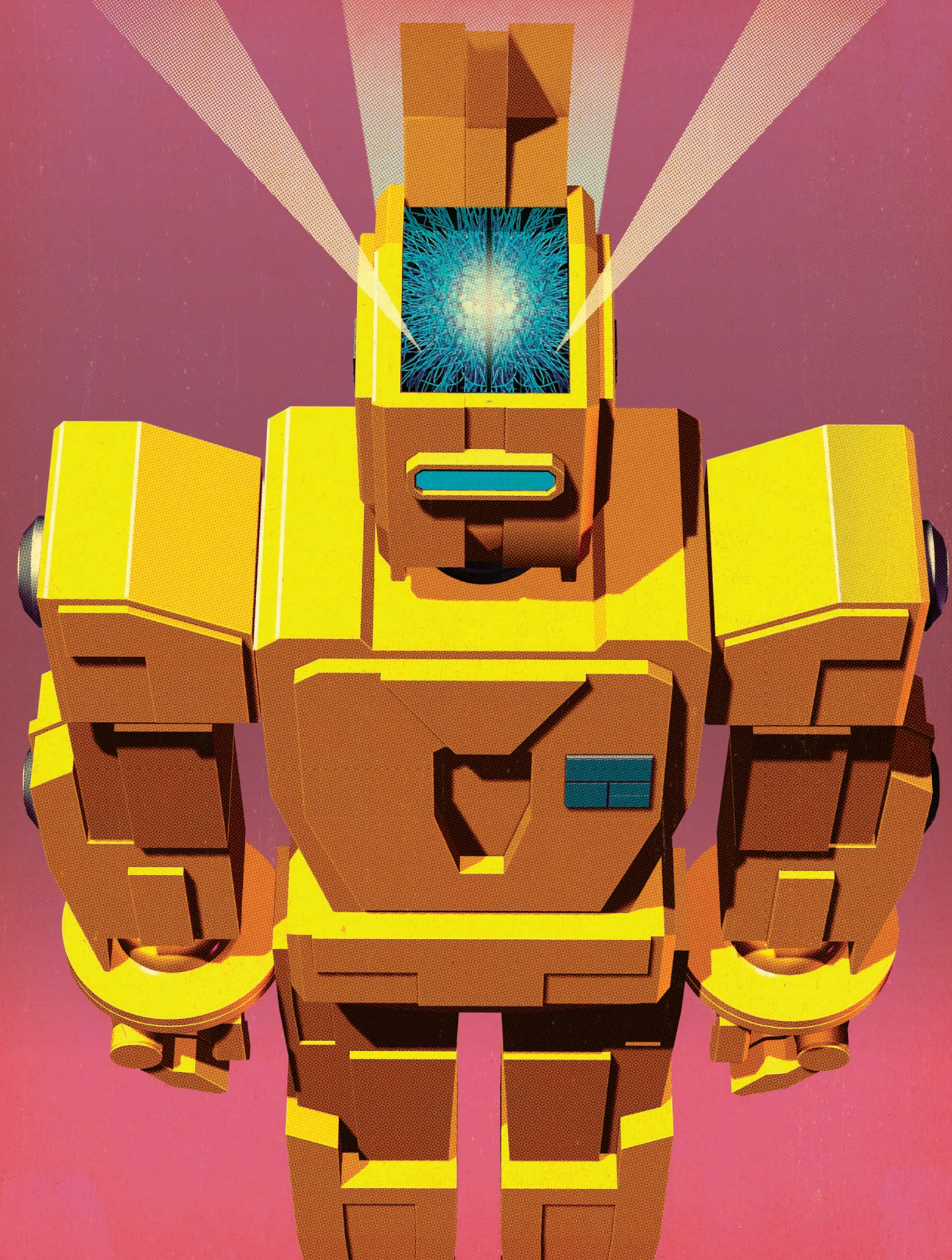


THE RISE OF AI

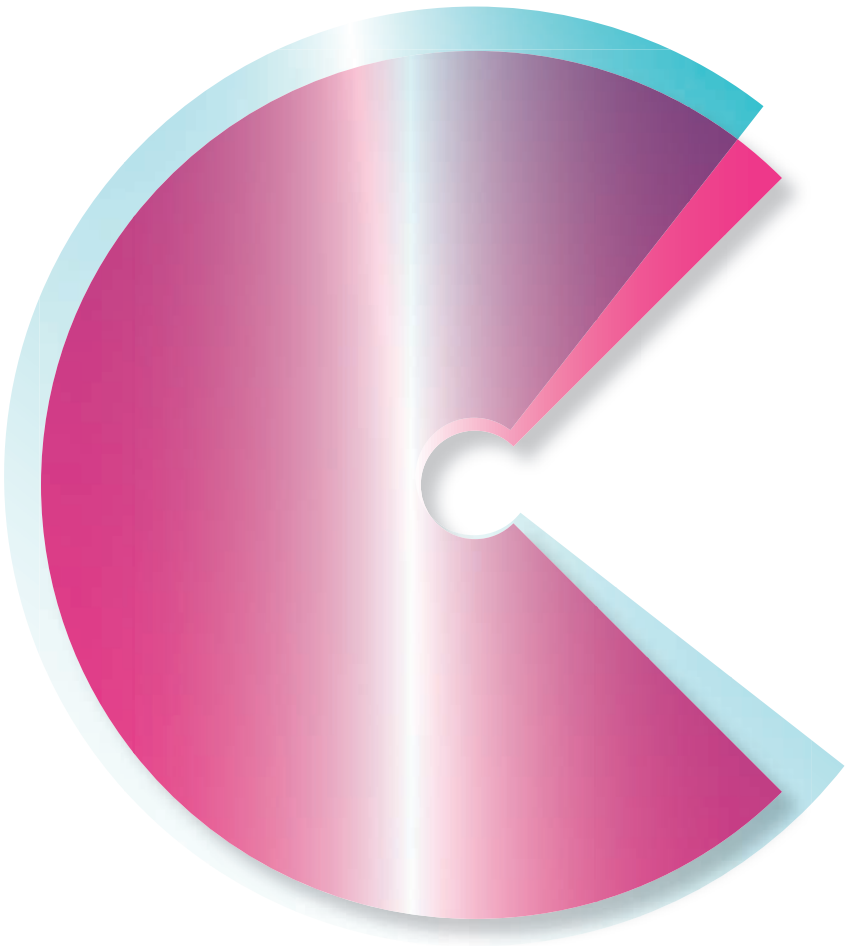
AI SPECIAL REPORT

The promise of driverless cars and software that teaches itself new skills has sparked a revival of artificial intelligence—and, with it, fears that our machines may one day turn against us

Illustrations by Tavis Coburn

COMPUTER SCIENCE

MACHINES WHO LEARN



After decades of disappointment, artificial intelligence is finally catching up to its early promise, thanks to a powerful technique called deep learning

By Yoshua Bengio

COMPUTERS GENERATED A GREAT DEAL of excitement in the 1950s when they began to beat humans at checkers and to prove math theorems. In the 1960s the hope grew that scientists might soon be able to replicate the human brain in hardware and software and that "artificial intelligence" would soon match human performance on any task. In 1967 Marvin Minsky of the Massachusetts Institute of Technology, who died earlier this year, proclaimed that the challenge of AI would be solved within a generation.

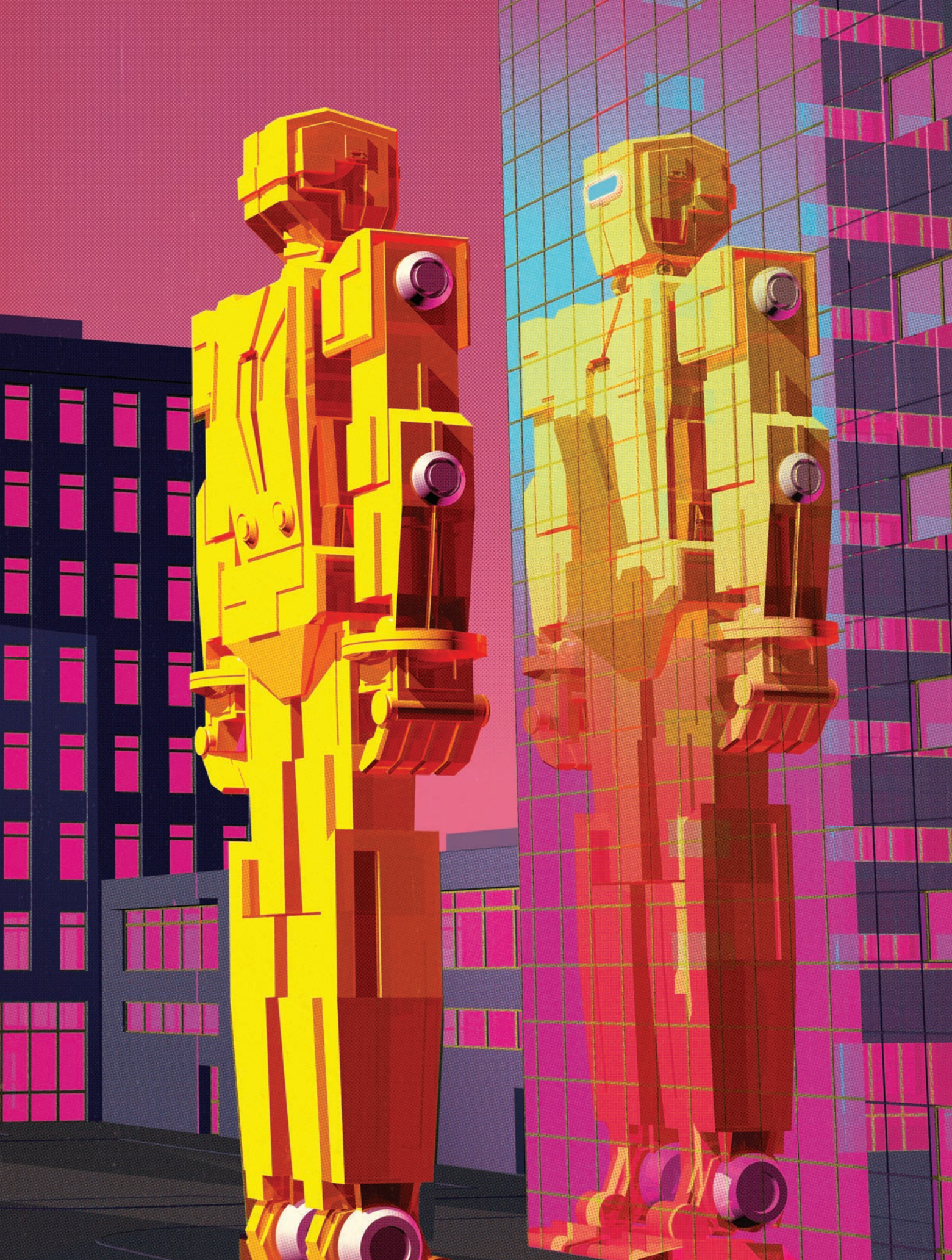
IN BRIEF

Artificial intelligence started as a field of serious study in the mid-1950s. At the time, investigators expected to emulate human intelligence within the span of an academic career.

Hopes were dashed when it became clear that the algorithms and computing power of that period were simply not up to the task. Some skeptics even wrote off the endeavor as pure hubris.

A revival took place during the past few years as software patterned roughly after networks of neurons in the brain demonstrated that AI's early promise might yet be realized.

Deep learning—a technique that uses complex neural networks—has the ability to learn abstract concepts and already approaches human-level performance on some tasks.



That optimism, of course, turned out to be premature. Software designed to help physicians make better diagnoses and networks modeled after the human brain for recognizing the contents of photographs failed to live up to their initial hype. The algorithms of those early years lacked sophistication and needed more data than were available at the time. Computer processing was also too tepid to power machines that could perform the massive calculations needed to approximate something approaching the intricacies of human thought.

By the mid-2000s the dream of building machines with human-level intelligence had almost disappeared in the scientific community. At the time, even the term "AI" seemed to leave the domain of serious science. Scientists and writers describe the dashed hopes of the period from the 1970s until the mid-2000s as a series of "AI winters."

What a difference a decade makes. Beginning in 2005, AI's outlook changed spectacularly. That was when deep learning, an approach to building intelligent machines that drew inspiration from brain science, began to come into its own. In recent years deep learning has become a singular force propelling AI research forward. Major information technology companies are now pouring billions of dollars into its development.

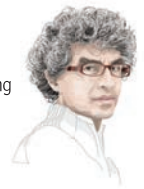
Deep learning refers to the simulation of networks of neurons that gradually "learn" to recognize images, understand speech or even make decisions on their own. The technique relies on so-called artificial neural networks—a core element of current AI research. Artificial neural networks do not mimic precisely how actual neurons work. Instead they are based on general mathematical principles that allow them to learn from examples to recognize people or objects in a photograph or to translate the world's major languages.

The technology of deep learning has transformed AI research, reviving lost ambitions for computer vision, speech recognition, natural-language processing and robotics. The first products rolled out in 2012 for understanding speech—you may be familiar with Google Now. And shortly afterward came applications for identifying the contents of an image, a feature now incorporated into the Google Photos search engine.

Anyone frustrated by clunky automated telephone menus can appreciate the dramatic advantages of using a better personal assistant on a smartphone. And for those who remember how poor object recognition was just a few years ago—software that might mistake an inanimate object for an animal—strides in computer vision have been incredible: we now have computers that, under certain conditions, can recognize a cat, a rock or faces in images almost as well as humans. AI software, in fact, has now become a familiar fixture in the lives of millions of smartphone users. Personally, I rarely type messages anymore. I often just speak to my phone, and sometimes it even answers back.

These advances have suddenly opened the door to further commercialization of the technology, and the excitement only continues to grow. Companies compete fiercely for talent, and Ph.D.s specializing in deep learning are a rare commodity that is in extremely high demand. Many university professors with expertise in this area—by some counts, the majority—have been pulled from academia to industry and furnished with well-appointed research facilities and ample compensation packages.

Working through the challenges of deep learning has led to stunning successes. The triumph of a neural network over top-ranked player Lee Se-dol at the game of Go received prominent headlines. Applications are already expanding to encompass other fields of human expertise—and it is not all games. A newly developed deep-learning algorithm is purported to diagnose heart failure from magnetic resonance imaging as well as a cardiologist.



Yoshua Bengio is a professor of computer science at the University of Montreal and one of the pioneers in developing the deep-learning methods that have sparked the current revival of artificial intelligence.

INTELLIGENCE, KNOWLEDGE AND LEARNING

WHY DID AI HIT so many roadblocks in previous decades? The reason is that most of the knowledge we have of the world around us is not formalized in written language as a set of explicit tasks—a necessity for writing any computer program. That is why we have not been able to directly program a computer to do many of the things that we humans do so easily—be it understanding speech, images or language or driving a car. Attempts to do so—organizing sets of facts in elaborate databases to imbue computers with a facsimile of intelligence—have met with scant success.

That is where deep learning comes in. It is part of the broader AI discipline known as machine learning, which is based on principles used to train intelligent computing systems—and to ultimately let machines teach themselves. One of these tenets relates to what a human or machine considers a "good" decision. For animals, evolutionary principles dictate decisions should be made that lead to behaviors that optimize chances of survival and reproduction. In human societies, a good decision might include social interactions that bring status or a sense of well-being. For a machine, such as a self-driving car, though, the quality of decision making depends on how closely the autonomous vehicle imitates the behaviors of competent human drivers.

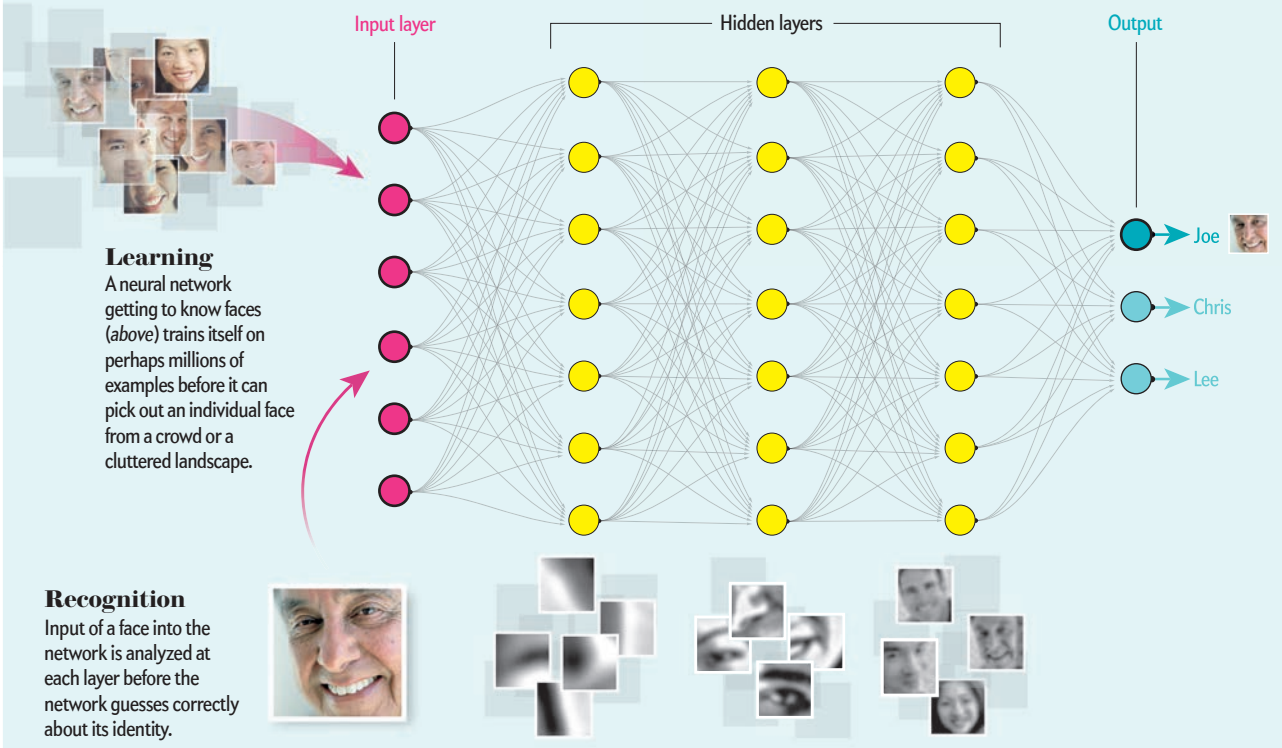
The knowledge needed to make a good decision in a particular context is not necessarily obvious in a way that can be translated into computer code. A mouse, for instance, has knowledge of its surroundings and an innate sense of where to sniff and how to move its legs, find food or mates, and avoid predators. No programmer would be capable of specifying a step-by-step set of instructions to produce these behaviors. Yet that knowledge is encoded in the rodent's brain.

Before creating computers that can train themselves, computer scientists needed to answer such fundamental questions as how humans acquire knowledge. Some knowledge is innate, but most is learned from experience. What we know intuitively cannot be turned into a clear sequence of steps for a computer to execute but can often be learned from examples and practice. Since the 1950s researchers have looked for and tried to refine general principles that allow animals or humans—or even machines, for that matter—to acquire knowledge through experience. Machine learning aims to establish procedures, called learning algorithms, that allow a machine to learn from examples presented to it.

Brainy Networks That Only Get Smarter

Connections from one neuron to the next in the brain's cortex have inspired the creation of algorithms that mimic these intricate links. A neural network can be trained to recognize a face by first training on countless images. Once it has "learned" to categorize a face (versus a hand, for instance) and to detect individual faces, the network uses that knowledge to identify faces it has seen before, even if the image of the person is slightly different from the one it was trained on.

To recognize a face, the network sets about the task of analyzing the individual pixels of an image presented to it at the input layer. Then, at the next layer, it chooses geometric shapes distinctive to a particular face. Moving up the hierarchy, a middle layer detects eyes, a mouth and other features before a composite full-face image is discerned at a higher layer. At the output layer, the network makes a "guess" about whether the face is that of Joe or rather that of Chris or Lee.



Learning
A neural network getting to know faces (*above*) trains itself on perhaps millions of examples before it can pick out an individual face from a crowd or a cluttered landscape.

Recognition
Input of a face into the network is analyzed at each layer before the network guesses correctly about its identity.

Each layer identifies progressively more complex features

PUNCHSTOCK (*faces*)

The science of machine learning is largely experimental because no universal learning algorithm exists—none can enable the computer to learn every task it is given well. Any knowledge-acquisition algorithm needs to be tested on learning tasks and data specific to the situation at hand, whether it is recognizing a sunset or translating English into Urdu. There is no way to prove that it will be consistently better across the board for any given situation than all other algorithms.

AI researchers have fashioned a formal mathematical description of this principle—the "no free lunch" theorem—that demonstrates that no algorithm exists to address every real-world learning situation. Yet human behavior apparently contradicts this theorem. We appear to hold in our head fairly general learning abilities that allow us to master a multitude of tasks for which evolution did not prepare our ancestors: playing chess, building bridges or doing research in AI.

These capabilities suggest that human intelligence exploits general assumptions about the world that might serve as inspiration for creating machines with a form of general intelligence. For just this reason, developers of artificial neural networks have adopted the brain as a rough model for designing intelligent systems.

The brain's main units of computation are cells called neurons. Each neuron sends a signal to other neurons through tiny gaps between the cells known as synaptic clefts. The propensity of a neuron to send a signal across the gap—and the amplitude of that signal—is referred to as synaptic strength. As a neuron "learns," its synaptic strength grows, and it is more likely, when stimulated by an electrical impulse, to send messages along to its neighbors.

Brain science influenced the emergence of artificial neural networks that used software or hardware to create virtual neurons. Early researchers in this subfield of AI, known as connectionism, postulated that neural networks would be able to learn

Graphic by Jen Christiansen

complex tasks by gradually altering the connections among neurons, so that patterns of neural activity would capture the content of its input, such as an image or a snippet of dialogue. As these networks would receive more examples, the learning process would continue by changing synaptic strengths among the connected neurons to achieve more accurate representations of, say, images of a sunset.

LESSONS ABOUT SUNSETS

THE CURRENT GENERATION of neural networks extends the pioneering work of connectionism. The networks gradually change numerical values for each synaptic connection, values representing the strength of that connection and thus how likely a neuron is to transmit a signal along to another neuron. An algorithm used by deep-learning networks changes these values ever so slightly each time it observes a new image. The values inch steadily closer toward ones that allow the neural network to make better predictions about the image's content.

For best results, current learning algorithms require close involvement by a human. Most of these algorithms use supervised learning in which each training example is accompanied by a human-crafted label about what is being learned—a picture of a sunset, say, is associated with a caption that says "sunset." In this instance, the goal of the supervised learning algorithm is to take a photograph as the input and produce, as an output, the name of a key object in the image. The mathematical process of transforming an input to an output is called a function. The numerical values, such as synaptic strengths, that produce this function correspond to a solution to the learning task.

Learning by rote to produce correct answers would be easy but somewhat useless. We want to teach the algorithm what a sunset is but then to have it recognize an image of any sunset, even one it has not been trained on. The ability to discern any sunset—in other words, to generalize learning beyond specific examples—is the main goal of any machine-learning algorithm. In fact, the quality of training of any network is evaluated by testing it using examples not previously seen. The difficulty of generalizing correctly to a new example arises because there is an almost infinite set of possible variations that still correspond to any category, such as a sunset.

To succeed in generalizing from having observed a multitude of examples, the learning algorithm used in deep-learning networks needs more than just the examples themselves. It also relies on hypotheses about the data and assumptions about what a possible solution to a particular problem might be. A typical hypothesis built into the software might postulate that if data inputs for a particular function are similar, the outputs should not radically change—altering a few pixels in an image of a cat should not usually transform the animal into a dog.

One type of neural network that incorporates hypotheses about images is called a convolutional neural network; it has become a key technology that has fueled the revival of AI. Convolutional neural networks used in deep learning have many layers of neurons organized in such a way as to make the output less sensitive to the main object in an image changing, such as when its position is moved slightly—a well-trained network may be able to recognize a face from different angles in separate photographs. The design of a convolutional network draws its inspiration from the multilayered structure of the visual cortex—the part of our brain that receives input from the eyes. The many layers of virtual neurons in a convolutional neural network are what makes a network "deep" and thus better able to learn about the world around it.

GOING DEEP

ON A PRACTICAL LEVEL, the advances that enabled deep learning came from specific innovations that emerged about 10 years ago, when interest in AI and neural networks had reached its lowest point in decades. A Canadian organization funded by the government and private donors, the Canadian Institute for Advanced Research (CIFAR), helped to rekindle the flame by sponsoring a program led by Geoffrey Hinton of the University of Toronto. The program also included Yann LeCun of New York University, Andrew Ng of Stanford University, Bruno Olshausen of the University of California, Berkeley, me and several others. Back then, negative attitudes toward this line of research made it difficult to publish and even to convince graduate students to work in this area, but a few of us had the strong sense that it was important to move ahead.

> The strong comeback for AI after a long and extended hiatus provides a lesson in the sociology of science, underscoring the need to put forward ideas that challenge the technological status quo.

Skepticism about neural networks at that time stemmed, in part, from the belief that training them was hopeless because of the challenges involved in optimizing how they behave. Optimization is a branch of mathematics that tries to find the configuration of a set of parameters to reach a mathematical objective. The parameters, in this case, are called synaptic weights and represent how strong a signal is being sent from one neuron to another.

The objective is to make predictions with the minimum number of errors. When the relation between parameters and an objective is simple enough—more precisely when the objective is a convex function of the parameters—the parameters can be gradually adjusted. This continues until they get as close as

possible to the values that produce the best possible choice, known as a global minimum—which corresponds to the lowest possible average prediction error made by the network.

In general, however, training a neural network is not so simple—and requires what is called a nonconvex optimization. This type of optimization poses a much greater challenge—and many researchers believed that the hurdle was insurmountable. The learning algorithm can get stuck in what is called a local minimum, in which it is unable to reduce the prediction error of the neural network by adjusting parameters slightly.

Only in the past year was the myth dispelled that neural networks were hard to train because of local minima. We found in our research that when a neural network is sufficiently large, the local minima problem is greatly reduced. Most local minima actually correspond to having learned knowledge at a level that almost matches the optimal value of the global minimum.

Although the theoretical problems of optimization could, in theory, be solved, building large networks with more than two or three layers had often failed. Beginning in 2005, CIFAR-supported efforts achieved breakthroughs that overcame these barriers. In 2006 we managed to train deeper neural networks, using a technique that proceeded layer by layer.

Later, in 2011, we found a better way to train even deeper networks—ones with more layers of virtual neurons—by altering the computations performed by each of these processing units, making them more like what biological neurons actually compute. We also discovered that injecting random noise into the signals transmitted among neurons during training, similar to what happens in the brain, made them better able to learn to correctly identify an image or sound.

Two crucial factors aided the success of deep-learning techniques. An immediate 10-fold increase in computing speed, thanks to the graphics-processing units initially designed for video games, allowed larger networks to be trained in a reasonable amount of time. Also fueling deep learning's growth was the availability of huge labeled data sets for which a learning algorithm can identify the correct answer—"cat," for example, when inspecting an image in which a cat is just one element.

Another reason for deep learning's recent success is its ability to learn to perform a sequence of computations that construct or analyze, step by step, an image, a sound or other data. The depth of the network is the number of such steps. Many visual- or auditory-recognition tasks in which AI excels require the many layers of a deep network. In recent theoretical and experimental studies, in fact, we have shown that carrying out some of these mathematical operations cannot be accomplished efficiently without sufficiently deep networks.

Each layer in a deep neural network transforms its input and produces an output that is sent to the next layer. The network represents more abstract concepts at its deeper layers [*see box on page 49*], which are more remote from the initial raw sensory input. Experiments show that artificial neurons in deeper layers in the network tend to correspond to more abstract semantic concepts: a visual object such as a desk, for instance. Recognition of the image of the desk might emerge from the processing of neurons at a deeper layer even though the concept of "desk" was not among the category labels on which the network was trained. And the concept of a desk might itself only be an intermediate step toward creating a still more abstract concept at a still higher layer that might be categorized by the network as an "office scene."

BEYOND PATTERN RECOGNITION

UNTIL RECENTLY, artificial neural networks distinguished themselves in large part for their ability to carry out tasks such as recognizing patterns in static images. But another type of neural network is also making its mark—specifically, for events that unfold over time. Recurrent neural networks have demonstrated the capacity to correctly perform a sequence of computations, typically for speech, video and other data. Sequential data are made up of units—whether a phoneme or a whole word—that follow one another sequentially. The way recurrent neural networks process their inputs bears a resemblance to how the brain works. Signals that course among neurons change constantly as inputs from the senses are processed. This internal neural state changes in a way that depends on the current input to the brain from its surroundings before issuing a sequence of commands that result in body movements directed at achieving a specific goal.

Recurrent networks can predict what the next word in a sentence will be, and this can be used to generate new sequences of words, one at a time. They can also take on more sophisticated tasks. After "reading" all the words in a sentence, the network can guess at the meaning of the entire sentence. A separate recurrent network can then use the semantic processing of the first network to translate the sentence into another language.

Research on recurrent neural networks had its own lull in the late 1990s and early 2000s. My theoretical work suggested that they would run into difficulty learning to retrieve information from the far past—the earliest elements in the sequence being processed. Think of trying to recite the words from the first sentences of a book verbatim when you have just reached the last page. But several advances have lessened some of these problems by enabling such networks to learn to store information so that it persists for an extended time. The neural networks can use a computer's temporary memory to process multiple, dispersed pieces of information, such as ideas contained in different sentences spread across a document.

The strong comeback for deep neural networks after the long AI winter is not just a technological triumph. It also provides a lesson in the sociology of science. In particular, it underscores the need to support ideas that challenge the technological status quo and to encourage a diverse research portfolio that backs disciplines that temporarily fall out of favor. SA

MORE TO EXPLORE

ImageNet Classification with Deep Convolutional Neural Networks. Alex Krizhevsky et al. Presented at the 26th Annual Conference on Neural Information Processing Systems (NIPS 2012), Stateline, Nev., December 3–8, 2012.

Representation Learning: A Review and New Perspectives. Y. Bengio et al. in *IEEE Transactions on Pattern Analysis and Machine Intelligence,* Vol. 35, No. 8, pages 1798–1828; August 2013.

Deep Learning. Yann LeCun et al. in *Nature,* Vol. 521, pages 436–444; May 28, 2015.

FROM OUR ARCHIVES

When Computers Surpass Us. Christof Koch; Consciousness Redux, *Scientific American Mind,* September/October 2015.

TRANSPORTATION

THE TRUTH ABOUT "SELF-DRIVING" CARS

AI SPECIAL REPORT

They are coming, but not the way you may have been led to think

By Steven E. Shladover

Steven E. Shladover helped to create the California Partners for Advanced Transportation Technology (PATH) program at the Institute of Transportation Studies at the University of California, Berkeley, in the 1980s. He is a mechanical engineer by training, with bachelor's, master's and doctoral degrees from the Massachusetts Institute of Technology.

SOON ELECTRONIC CHAUFFEURS WILL TAKE US WHEREVER WE WANT to go, whenever we want, in complete safety—as long as we do not need to make any left turns across traffic. Changing road surfaces are a problem, too. So are snow and ice. It will be crucial to avoid traffic cops, crossing guards and emergency vehicles. And in an urban environment where pedestrians are likely to run out in front of the car, we should probably just walk or take the subway.

* *

All these simple, everyday encounters for human drivers pose enormous problems for computers that will take time, money and effort to solve. Yet much of the public is becoming convinced that fully automated vehicles are just around the corner.

What created this disconnect? Part of the problem is terminology. The popular media applies the descriptors "autonomous," "driverless" and "self-driving" indiscriminately to technologies that are very different from one another, blurring important distinctions. And the automotive industry has not helped clarify matters. Marketers working for vehicle manufacturers, equipment suppliers and technology companies carefully compose publicity materials to support a wide range of interpretations about the amount of driving their products automate. Journalists who cover the field have an incentive to adopt the most optimistic forecasts—they are simply more exciting. The result of this feedback loop is a spiral of increasingly unrealistic expectations.

This confusion is unfortunate because automated driving is coming, and it could save lives, reduce pollution and conserve fuel. But it will not happen in quite the way you have been told.

DEFINING AUTOMATED DRIVING

DRIVING IS A MUCH MORE COMPLEX ACTIVITY than most people appreciate. It involves a broad range of skills and actions, some of which are easier to automate than others. Maintaining speed on an open road is simple, which is why conventional cruise-control systems have been doing it automatically for decades. As technology has advanced, engineers have been able to automate additional driving subtasks. Widely available adaptive cruise-control systems now maintain proper speed and spacing behind other vehicles. Lane-keeping systems, such as those in new models from Mercedes-Benz and Infiniti, use cameras, sensors and steering control to keep a vehicle centered in its lane. Cars are pretty smart these days. Yet it is an enormous leap from such systems to fully automated driving.

A five-level taxonomy defined by SAE International (formerly

IN BRIEF

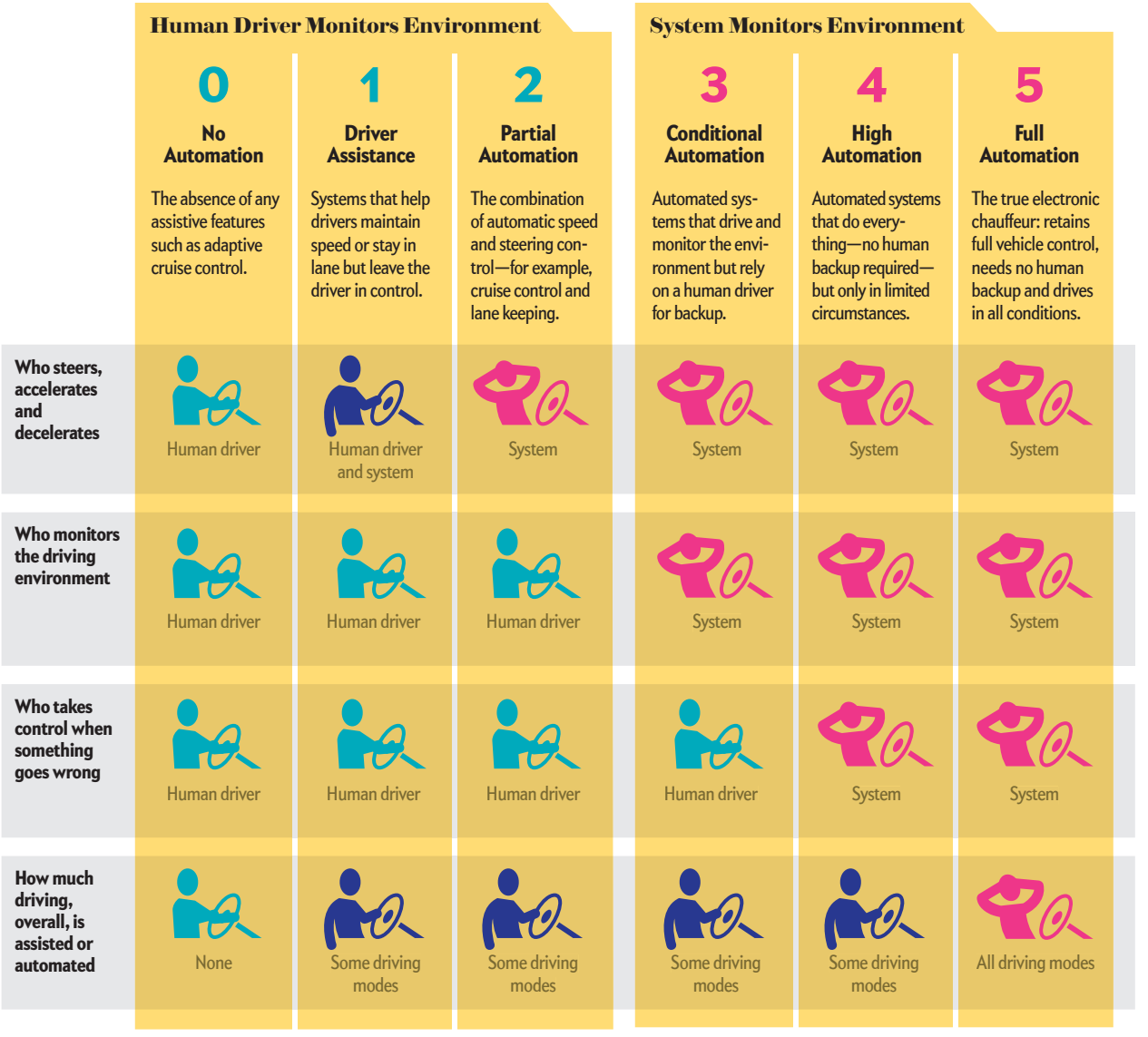
The auto industry and the press have oversold the automated car. Simple road encounters pose huge challenges for computers, and robotic chauffeurs remain decades away.

Automated driving systems that rely on humans for backup are particularly problematic. Yet in the next decade we will see automatic-driving systems that are limited to specific conditions and applications.

Automatic parking valets, low-speed campus shuttles, closely spaced platoons of heavy trucks and automatic freeway-control systems for use in dedicated lanes are all feasible and perhaps inevitable.

The Ladder of Automation

The automotive industry and the media have made a mess of the terminology used to talk about automated driverless systems. The terms "autonomous," "driverless" and "self-driving" obscure more than they illuminate. To clear things up, SAE International wrote definitions, paraphrased here, for different levels of automation and arranged them on a ladder of decreasing reliance on the driver. The hierarchy reveals some surprises. For example, level-four automation is potentially more tractable than level three. Level-five automated systems—electronic chauffeurs that can handle any driving condition with no human input—are decades away.



SOURCE: SAE INTERNATIONAL (WWW.SAE.ORG/MISC/PDFS/AUTOMATED_DRIVING.PDF); FOR DOCUMENTATION, INCLUDING ORIGINAL DEFINITIONS IN THEIR ENTIRETY, SEE *TAXONOMY AND DEFINITIONS FOR TERMS RELATED TO ON-ROAD MOTOR VEHICLE AUTOMATED DRIVING SYSTEMS*. SAE INTERNATIONAL, JANUARY 2014

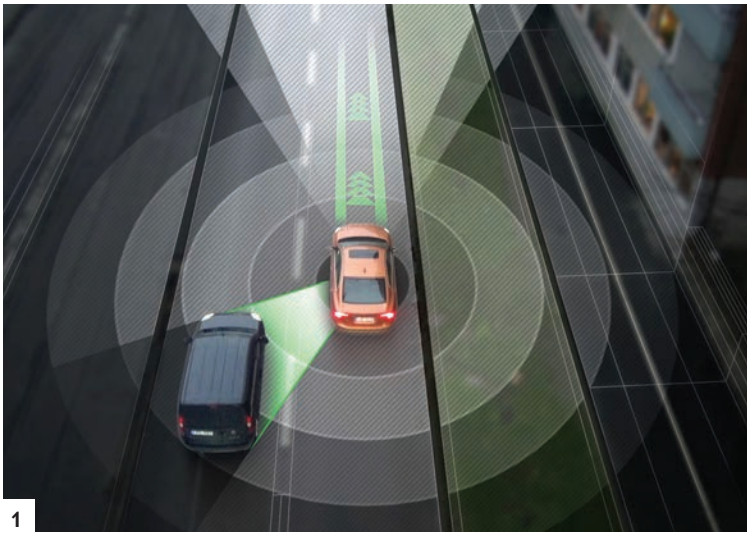
the Society of Automotive Engineers) is useful for clarifying our thinking about automated driving. The first three rungs on this ladder of increasing automation (excluding level zero, for no automation) are occupied by technologies that rely on humans for emergency backup. Adaptive cruise control, lane-keeping systems, and the like belong to level one. Level-two systems combine the functions of level-one technologies—the lateral and longitudinal controls of lane-keeping and adaptive cruise-control systems, for example—to automate more complex driving tasks. This is as far as commercially available vehicle automation goes today. Level-three systems would allow drivers to turn on autopilot in specific scenarios, such as freeway traffic jams.

The next two levels are profoundly different in that they operate entirely without human assistance. Level-four (high-auto-

Illustrations by Nigel Holmes (icons)

mation) systems would handle all driving subtasks, but they would operate only in strictly defined scenarios—in enclosed parking garages, for example, or in dedicated lanes on the freeway. At the top of the ladder is level five—the fully automated car. Presumably, this is what many people have in mind when they hear someone such as Nissan CEO Carlos Ghosn confidently proclaim that automated cars will be on the road by 2020.

The truth is that no one expects level-five automation systems to be on the market by then. In all likelihood, they are a long way off. Level-three systems might be just as remote. But level four? Look for it within the next decade. To understand this confusing state of affairs, we have to talk about software.



1

2

NEXT YEAR Volvo Cars will field-test 100 vehicles equipped with systems that automate driving on special stretches of freeway (*1* and *2*). Volvos have also been used in European road-train tests (*3*).

SOFTWARE NIGHTMARE

DESPITE THE POPULAR PERCEPTION, human drivers are remarkably capable of avoiding serious crashes. Based on the total U.S. traffic safety statistics for 2011, fatal crashes occurred about once for every 3.3 million hours of driving; crashes that resulted in injury happened approximately once for every 64,000 hours of driving. These numbers set an important safety target for automated driving systems, which should, at minimum, be no less safe than human drivers. Reaching this level of reliability will require vastly more development than automation enthusiasts want to admit.

Think about how often your laptop freezes up. If that software were responsible for driving a car, the "blue screen of death" would become more than a figure of speech. A delayed software response of as little as one tenth of a second is likely to be hazardous in traffic. Software for automated driving must therefore be designed and developed to dramatically different standards from anything currently found in consumer devices.

Achieving these standards will be profoundly difficult and require basic breakthroughs in software engineering and signal processing. Engineers need new methods for designing software that can be proved correct and safe even in complex and rapidly changing conditions. Formal methods for analyzing every possible failure mode for a piece of code before it is written exist—think of them as mathematical proofs for computer programs—but only for very simple applications. Scientists are only beginning to think about how to scale up these kinds of tests to validate the incredibly complex code required to control a fully automated vehicle.

Once that code has been written, software engineers will need new methods for debugging and verifying it. Existing methods are too cumbersome and costly for the job. To put this in perspective, consider that half of the cost of a new commercial or military aircraft goes toward software verification and validation. The software on aircraft is actually much *less* complex than what will be needed for automated road vehicles. An engineer can design an aircraft autopilot system knowing that it will rarely, if ever, have to deal with more than one or two other aircraft in its vicinity. It does not need to know the velocity and location of those aircraft with incredible precision, because they are far enough apart that they have time to act. Decisions must be made on the order of tens of seconds. An automated road vehicle will have to track dozens of other vehicles and obstacles and make decisions within fractions of a second. The code required will be orders of magnitude more complex than what it takes to fly an airplane.

Once the code is validated, manufacturers will need ways to "prove" the safety of a complete automated driving system to the satisfaction of company risk-management officers, insurance firms, safety advocates, regulators and, of course, potential customers. The kind of formal "acceptance tests" used today are completely impractical for this purpose. Testers would have to put hundreds of millions, if not billions, of miles on a vehicle to ensure that they have subjected it in a statistically significant way to the dangerous scenarios it will encounter when it is regularly used by thousands of customers. People have started to think about solutions to this problem—the German government and industry have launched a multimillion-dollar project with that goal—but those efforts have just begun.

The code that will control the vehicle—the brain, so to speak—is not the only thing that must be subjected to scrutiny. The sensors that provide that brain with the data it will use to make decisions must be subjected to equal scrutiny. Engineers must develop new sensor-signal processing and data-fusion algorithms that can discriminate between benign and hazardous objects in a vehicle's path with nearly zero false negatives (hazardous objects

COURTESY OF VOLVO CAR GROUP (1-3)

3

that were not identified) and extremely low false positives (benign objects that were misclassified, leading to inappropriate responses from vehicles, such as swerving or hard braking).

Engineers cannot resort to the kind of brute-force redundancy used in commercial aircraft systems to achieve these goals because an automated car is a consumer product: it must be affordable for the general public. Turning to artificial intelligence is not an obvious solution, either. Some people have suggested that machine-learning systems could enable automated driving systems to study millions of hours of driving data and then learn throughout the course of their life cycle. But machine learning introduces its own problems because it is nondeterministic. Two identical vehicles can roll off the assembly line, but after a year of encountering different traffic situations, their automation systems will behave very differently.

LEVEL-FOUR FUTURE

I USED TO TELL PEOPLE that level-five fully automated driving systems would not become feasible until after 2040. Somewhere along the way people started quoting me as saying level five would arrive *in* 2040. Now I say that fully automated vehicles capable of driving in every situation will not be here until 2075. Could it happen sooner than that? Certainly. But not by much.

The prospects for level-three automation are clouded, too, because of the very real problem of recapturing the attention, in an emergency, of a driver who has zoned out while watching the scenery go by or, worse, who has fallen asleep. I have heard representatives from some automakers say that this is such a hard problem that they simply will not attempt level three. Outside of traffic-jam assistants that take over in stop-and-go traffic, where speeds are so low that a worst-case collision would be a fender bender, it is conceivable that level-three automation will never happen.

And yet we will see highly automated cars soon, probably within the coming decade. Nearly every big automaker and many information technology companies are devoting serious resources to level-four automation: fully automated driving, restricted to specific environments, that does not rely on a fallible human for backup. When you limit the situations in which automated vehicle systems must operate, you greatly increase their feasibility. (Automated people movers have been operating in big airports for years—but they are on totally segregated tracks.)

In all probability, the next 10 years will bring automated valet-parking systems that will allow drivers to drop their cars at the entrance of a suitably equipped garage that excludes pedestrians and nonautomated vehicles. An onboard automation system will communicate with sensors placed throughout the garage to find out which parking spots are available and navigate to them. Because there will be no need to open the doors, parking spaces can be narrower than they are today, so more cars will be able to fit in garages in areas where space is expensive.

In urban pedestrian zones, business parks, university campuses and other places where high-speed vehicles can be excluded, low-speed passenger shuttles will operate without drivers. In such environments, limited-capability sensors should be adequate to detect pedestrians and bicyclists, and if a sensor detects a false positive and brakes unnecessarily, it will not harm anyone (although it will annoy the people in the vehicle). The CityMobil2 project of the European Commission has been demonstrating such technologies for several years, and its final demonstration is scheduled for this summer.

Segregated bus ways and truck-only lanes will soon enable commercial vehicles to operate at higher levels of automation. Physically segregating these vehicles from other users will greatly simplify threat detection and response systems. Eventually driverless trucks and buses will be able to follow a human-driven lead vehicle in fuel-saving platoons. Researchers worldwide, including the California Partners for Advanced Transportation Technology (PATH) program at the University of California, Berkeley, Japan's Energy ITS project, and the KONVOI and SARTRE projects in Europe, have already tested prototype bus- and truck-platoon systems.

Yet the most widespread implementation of level-four automation within the next decade will probably be automated freeway systems for personal passenger vehicles. These systems will permit automobiles to drive themselves under certain conditions on designated sections of freeway. The vehicles will have redundant components and subsystems so that if something goes wrong, they can "limp home" without human guidance. They will probably be restricted to fair weather on stretches of freeway that have been mapped in detail, down to the signage and lane markings. These sections of road might even have "safe harbor" locations where vehicles can go when they have problems. Most major vehicle manufacturers are hard at work developing these systems, and next year Volvo Cars plans to conduct a public field test of such capabilities with 100 prototype vehicles in Gothenburg, Sweden.

These scenarios might not sound as futuristic as having your own personal electronic chauffeur, but they have the benefit of being possible—even inevitable—and soon. SA

MORE TO EXPLORE

Technical Challenges for Fully Automated Driving Systems. Steven Shladover. Presented at the 21st World Congress on Intelligent Transport Systems, Detroit, Mich., September 7–11, 2014.

Towards Road Transport Automation: Opportunities in Public-Private Collaboration. Summary of the Third EU-U.S. Transportation Research Symposium, Washington, D.C., April 14–15, 2015. National Academies of Sciences, Engineering, and Medicine, 2015.

Summary of definitions for SAE International's 2014 report *Taxonomy and Definitions for Terms Related to On-Road Motor Vehicle Automated Driving Systems:* www.sae.org/misc/pdfs/automated_driving.pdf

FROM OUR ARCHIVES

Driving toward Crashless Cars. Steven Ashley; December 2008.

Stuart Russell is a professor of computer science at the University of California, Berkeley, and an expert on artificial intelligence.

COMMENTARY

SHOULD WE FEAR SUPERSMART ROBOTS?

If we're not careful, we could find ourselves at odds with determined, intelligent machines whose objectives conflict with our own

By Stuart Russell

IT IS HARD TO ESCAPE THE NAGGING SUSPICION THAT CREATING MACHINES smarter than ourselves *might* be a problem. After all, if gorillas had accidentally created humans way back when, the now endangered primates probably would be wishing they had not done so. But *why,* specifically, is advanced artificial intelligence a problem?

Hollywood's theory that spontaneously evil machine consciousness will drive armies of killer robots is just silly. The real problem relates to the possibility that AI may become incredibly good at achieving something other than what we really want. In 1960 legendary mathematician Norbert Wiener, who founded the field of cybernetics, put it this way: "If we use, to achieve our purposes, a mechanical agency with whose operation we cannot efficiently interfere…, we had better be quite sure that the purpose put into the machine is the purpose which we really desire."

A machine with a specific purpose has another property, one that we usually associate with living things: a wish to preserve its own existence. For the machine, this trait is not innate, nor is it something introduced by humans; it is a logical consequence of the simple fact that the machine cannot achieve its original purpose if it is dead. So if we send out a robot with the sole directive of fetching coffee, it will have a strong incentive to ensure success by disabling its own off switch or even exterminating anyone who might interfere with its mission. If we are not careful, then, we could face a kind of global chess match against very determined, superintelligent machines whose objectives conflict with our own, with the real world as the chessboard.

The prospect of entering into and losing such a match should concentrate the minds of computer scientists. Some researchers argue that we can seal the machines inside a kind of firewall, using them to answer difficult questions but never allowing them to affect the real world. (Of course, this means giving up on superintelligent robots!) Unfortunately, that plan seems unlikely to work: we have yet to invent a firewall that is secure against ordinary humans, let alone superintelligent machines.

Can we instead tackle Wiener's warning head-on? Can we de-

sign AI systems whose goals do not conflict with ours so that we are sure to be happy with the way they behave? This is far from easy—after all, stories with a genie and three wishes often end with a third wish to undo the first two—but I believe it is possible if we follow three core principles in designing intelligent systems:

The machine's purpose must be to maximize the realization of human values. In particular, the machine has no purpose of its own and no innate desire to protect itself.

The machine must be initially uncertain about what those human values are. This turns out to be crucial, and in a way it sidesteps Wiener's problem. The machine may learn more about human values as it goes along, of course, but it may never achieve complete certainty.

The machine must be able to learn about human values by observing the choices that we humans make.

The first two principles may seem counterintuitive, but together they avoid the problem of a robot having a strong incentive to disable its own off switch. The robot is sure it wants to maximize human values, but it also does not know exactly what those are. Now the robot actually *benefits* from being switched off because it understands that the human will press the off switch to prevent the robot from doing something counter to human values. Thus, the robot has a positive incentive to keep the off switch intact—and this incentive derives directly from its uncertainty about human values.

The third principle borrows from a subdiscipline of AI called inverse reinforcement learning (IRL), which is specifically concerned with learning the values of some entity—whether a human, canine or cockroach—by observing its behavior. By watching a typical human's morning routine, the robot learns about the value of coffee to humans. The field is in its infancy, but already some practical algorithms exist that demonstrate its potential in designing smart machines.

As IRL evolves, it must find ways to cope with the fact that humans are irrational, inconsistent and weak-willed and have limited computational powers, so their actions do not always reflect their values. Also, humans exhibit diverse sets of values, which means that robots must be sensitive to potential conflicts and trade-offs among people. And some humans are just plain evil and should be neither helped nor emulated.

Despite these difficulties, I believe it will be possible for machines to learn enough about human values that they will not pose a threat to our species. Besides directly observing human behavior, machines will be aided by having access to vast amounts of written and filmed information about people doing things (and others reacting). Designing algorithms that can understand this information is much easier than designing superintelligent machines. Also, there are strong economic incentives for robots—and their makers—to understand and acknowledge human values: if one poorly designed domestic robot cooks the cat for dinner, not realizing that its sentimental value outweighs its nutritional value, the domestic robot industry will be out of business.

Solving the safety problem well enough to move forward in AI seems to be feasible but not easy. There are probably decades in which to plan for the arrival of superintelligent machines. But the problem should not be dismissed out of hand, as it has been by some AI researchers. Some argue that humans and machines can coexist as long as they work in teams—yet that is not feasible unless machines share the goals of humans. Others say we can just "switch them off" as if superintelligent machines are too stupid to think of that possibility. Still others think that superintelligent AI will never happen. On September 11, 1933, renowned physicist Ernest Rutherford stated, with utter confidence, "Anyone who expects a source of power in the transformation of these atoms is talking moonshine." On September 12, 1933, physicist Leo Szilard invented the neutron-induced nuclear chain reaction. SA