The Three Breakthroughs That Have Finally Unleashed AI on the World

- BY KEVIN KELLY


CRAIG & KARL

A few months ago I made the trek to the sylvan campus of the IBM research labs in Yorktown Heights, New York, to catch an early glimpse of the fast-arriving, long-overdue future of artificial intelligence. This was the home of Watson, the electronic genius that conquered *Jeopardy!* in 2011. The original Watson is still here—it's about the size of a bedroom, with 10 upright, refrigerator-shaped machines forming the four walls. The tiny interior cavity gives technicians access to the jumble of wires and cables on the machines' backs. It is surprisingly warm inside, as if the cluster were alive.

Today's Watson is very different. It no longer exists solely within a wall of cabinets but is spread across a cloud of open-standard servers that run several hundred “instances” of the AI at once. Like all things cloudy, Watson is served to simultaneous customers anywhere in the world, who can access it using their phones, their desktops, or their own data servers. This kind of AI can be scaled up or down on demand. Because AI improves as people use it, Watson is always getting smarter; anything it learns in one instance can be immediately transferred to the others. And instead of one single program, it's an aggregation of diverse software engines—its logic-deduction engine and its language-parsing engine might operate on different code, on different chips, in different locations—all cleverly integrated into a unified stream of intelligence.

Consumers can tap into that always-on intelligence directly, but also through third-party apps that harness the power of this AI cloud. Like many parents of a bright mind, IBM would like Watson to pursue a medical career, so it should come as no surprise that one of the apps under development is a medical-diagnosis tool. Most of the previous attempts to make a diagnostic AI have been pathetic failures, but Watson really works. When, in plain English, I give it the symptoms of a disease I once contracted in India, it gives me a list of hunches, ranked from most to least probable. The most likely cause, it declares, is *Giardia*—the correct answer. This expertise isn't yet available to patients directly; IBM provides access to Watson's intelligence to partners, helping them develop user-friendly interfaces for subscribing doctors and hospitals. “I

believe something like Watson will soon be the world's best diagnostician—whether machine or human," says Alan Greene, chief medical officer of Scanadu, a startup that is building a diagnostic device inspired by the *Star Trek* medical tricorder and powered by a cloud AI. "At the rate AI technology is improving, a kid born today will rarely need to see a doctor to get a diagnosis by the time they are an adult."

AS AIS DEVELOP, WE MIGHT HAVE TO ENGINEER WAYS TO *PREVENT CONSCIOUSNESS* IN THEM—OUR MOST PREMIUM AI SERVICES WILL BE ADVERTISED AS CONSCIOUSNESS-FREE.

Medicine is only the beginning. All the major cloud companies, plus dozens of startups, are in a mad rush to launch a Watson-like cognitive service. According to quantitative analysis firm Quid, AI has attracted more than $17 billion in investments since 2009. Last year alone more than $2 billion was invested in 322 companies with AI-like technology. Facebook and Google have recruited researchers to join their in-house AI research teams. Yahoo, Intel, Dropbox, LinkedIn, Pinterest, and Twitter have all purchased AI companies since last year. Private investment in the AI sector has been expanding 62 percent a year on average for the past four years, a rate that is expected to continue.

Amid all this activity, a picture of our AI future is coming into view, and it is not the HAL 9000—a discrete machine animated by a charismatic (yet potentially homicidal) humanlike consciousness—or a Singularitan rapture of superintelligence. The AI on the horizon looks more like Amazon Web Services—cheap, reliable, industrial-grade digital smartness running behind everything, and almost invisible except when it blinks off. This common utility will serve you as much IQ as you want but no more than you need. Like all utilities, AI will be supremely boring, even as it transforms the Internet, the global economy, and civilization. It will enliven inert objects, much as electricity did more than a century ago. Everything that we formerly electrified we will now cognitize. This new utilitarian AI will also augment us individually as people (deepening our memory, speeding our recognition) and collectively as a species. There is almost nothing we can think of that cannot be made new, different, or interesting by infusing it with some extra IQ. In fact, the business plans of the next 10,000 startups are easy to forecast: *Take X and add AI*. This is a big deal, and now it's here.

Around 2002 I attended a small party for Google—before its IPO, when it only focused on search. I struck up a conversation with Larry Page, Google's brilliant cofounder, who became the company's CEO in 2011. “Larry, I still don't get it. There are so many search companies. Web search, for free? Where does that get you?” My unimaginative blindness is solid evidence that predicting is hard, especially about the future, but in my defense this was before Google had ramped up its ad-auction scheme to generate real income, long before YouTube or any other major acquisitions. I was not the only avid user of its search site who thought it would not last long. But Page's reply has always stuck with me: “Oh, we're really making an AI.”

I've thought a lot about that conversation over the past few years as Google has bought 14 AI and robotics companies. At first glance, you might think that Google is beefing up its AI portfolio to improve its search capabilities, since search contributes 80 percent of its revenue. But I think that's backward. Rather than use AI to make its search better, Google is using search to make its AI better. Every time you type a query, click on a search-generated link, or create a link on the web, you are training the Google AI. When you type “Easter Bunny” into the image search bar and then click on the most Easter Bunny-looking image, you are teaching the AI what an Easter bunny looks like. Each of the 12.1 billion queries that Google's 1.2 billion searchers conduct each day tutor the deep-learning AI over and over again. With another 10 years of steady improvements to its AI algorithms, plus a thousand-fold more data and 100 times more computing resources, Google will have an unrivaled AI. My prediction: By 2024, Google's main product will not be search but AI.

This is the point where it is entirely appropriate to be skeptical. For almost 60 years, AI researchers have predicted that AI is right around the corner, yet until a few years ago it seemed as stuck in the future as ever. There was even a term coined to describe this era of meager results and even more meager research funding: the AI winter. Has anything really changed?

Yes. Three recent breakthroughs have unleashed the long-awaited arrival of artificial intelligence:

1. Cheap parallel computation

Thinking is an inherently parallel process, billions of neurons firing simultaneously to create synchronous waves of cortical computation. To build a neural network—the primary architecture of AI software—also requires many different processes to take place simultaneously. Each node of a neural network loosely imitates a neuron in the brain—mutually interacting with its neighbors to make sense of the signals it receives. To recognize a spoken word, a program must be able to hear all the phonemes in relation to one another; to identify an image, it needs to see every pixel in the context of the pixels around it—both deeply parallel tasks. But until recently, the typical computer processor could only ping one thing at a time.

That began to change more than a decade ago, when a new kind of chip, called a graphics processing unit, or GPU, was devised for the intensely visual—and parallel—demands of videogames, in which millions of pixels had to be recalculated many times a second. That required a specialized parallel computing chip, which was added as a supplement to the PC motherboard. The parallel graphical chips worked, and gaming soared. By 2005, GPUs were being produced in such quantities that they became much cheaper. In 2009, Andrew Ng and a team at Stanford realized that GPU chips could run neural networks in parallel.

That discovery unlocked new possibilities for neural networks, which can include hundreds of millions of connections between their nodes. Traditional processors required several weeks to calculate all the cascading possibilities in a 100 million-parameter neural net. Ng found that a cluster of GPUs could accomplish the same thing in a day. Today neural nets running on GPUs are routinely used by cloud-enabled companies such as Facebook to identify your friends in photos or, in the case of Netflix, to make reliable recommendations for its more than 50 million subscribers.

2. Big Data

Every intelligence has to be taught. A human brain, which is genetically primed to categorize things, still needs to see a dozen examples before it can distinguish between cats and dogs. That's even more true for artificial minds. Even the best-programmed computer has to play at least a thousand games of chess before it gets good. Part of the AI breakthrough lies in the incredible avalanche of collected data about our world, which provides the schooling that AIs need.

Massive databases, self-tracking, web cookies, online footprints, terabytes of storage, decades of search results, Wikipedia, and the entire digital universe became the teachers making AI smart.

3. Better algorithms

Digital neural nets were invented in the 1950s, but it took decades for computer scientists to learn how to tame the astronomically huge combinatorial relationships between a million—or 100 million—neurons. The key was to organize neural nets into stacked layers. Take the relatively simple task of recognizing that a face is a face. When a group of bits in a neural net are found to trigger a pattern—the image of an eye, for instance—that result is moved up to another level in the neural net for further parsing. The next level might group two eyes together and pass that meaningful chunk onto another level of hierarchical structure that associates it with the pattern of a nose. It can take many millions of these nodes (each one producing a calculation feeding others around it), stacked up to 15 levels high, to recognize a human face. In 2006, Geoff Hinton, then at the University of Toronto, made a key tweak to this method, which he dubbed "deep learning." He was able to mathematically optimize results from each layer so that the learning accumulated faster as it proceeded up the stack of layers. Deep-learning algorithms accelerated enormously a few years later when they were ported to GPUs. The code of deep learning alone is insufficient to generate complex logical thinking, but it is an essential component of all current AIs, including IBM's Watson, Google's search engine, and Facebook's algorithms.

CRAIG & KARL

This perfect storm of parallel computation, bigger data, and deeper algorithms generated the 60-years-in-the-making overnight success of AI. And this convergence suggests that as long as these technological trends continue—and there's no reason to think they won't—AI will keep improving.

As it does, this cloud-based AI will become an increasingly ingrained part of our everyday life. But it will come at a price. Cloud computing obeys the law of increasing returns, sometimes called the network effect, which holds that the value of a network increases much faster as it grows bigger. The bigger the network, the more attractive it is to new users, which makes it even bigger, and thus more attractive, and so on. A cloud that serves AI will obey the same law. The more people who use an AI, the smarter it gets. The smarter it gets, the more people use it. The

more people that use it, the smarter it gets. Once a company enters this virtuous cycle, it tends to grow so big, so fast, that it overwhelms any upstart competitors. As a result, our AI future is likely to be ruled by an oligarchy of two or three large, general-purpose cloud-based commercial intelligences.

AI EVERYWHERE

Over the past five years, cheap computing, novel algorithms, and mountains of data have enabled new AI-based services that were previously the domain of sci-fi and academic white papers. —ROBERT MCMILLAN

GETTY

ALEMY

Self-Driving Car | Google has moved on from its initial goal of trying to index the entire Internet. Now it wants to index reality—part of its effort to perfect its self-driving car. Before the vehicle navigates a particular route, Google drivers scope out the course and then produce the most precise maps imaginable. That way the autonomous car knows what to expect and simply has to scan the environment with its roof-mounted lasers, cameras, and

radar systems to spot anything out of the ordinary. That's a much easier problem to solve than building a real-time map of the world.

ARIEL ZAMBELICH

Body Tracker | To turn the human body into a game controller, researchers working on Microsoft's Xbox Kinect had to deploy new machine-learning techniques. First, the device's infrared emitter and sensor create a 3-D image of a player's frame and analyze its different parts—shoulders, feet, hands. Then, using a method called decision forests, Kinect's AI system guesses the body's most likely next position. The result is a system that reads your movements in real time, without overwhelming the Xbox's memory.

Personal Photo Archivist |Matt Zeiler wants you to be able to find a snapshot as easily as you look up a phone number. His startup, Clarifai, is developing a new search technique to index the photos on your phone. While old-school image search looks for colors and lines, Clarifai's AI software understands corners and parallel lines, then can master higher-level concepts like wheels or cars as it studies more and more pictures.

Universal Translator | The Skype Translator, which will debut in beta by year's end, translates speech in real time, allowing anyone to talk naturally with anyone else. The AI software examines millions of translated sentences until it becomes superb at guessing how any given jumble of words will translate. For voice recognition, it breaks down samples of the spoken word, analyzing them until it achieves a sophisticated grasp of the ways sounds combine to form speech.

Smarter News Feed |Facebook hired one of the world's foremost deep-learning experts, Yann LeCun, to set up an AI lab last year. He's tasked with improving the social network's speech and image recognition software to make it more efficient at identifying, say, viral videos that you'll find funny or photos that you'll want to see—like your friends in a group snapshot.