MASTERING AI

A SURVIVAL GUIDE TO OUR SUPERPOWERED FUTURE

JEREMY KAHN

SIMON & SCHUSTER

NEW YORK LONDON TORONTO SYDNEY NEW DELHI

1230 Avenue of the Americas
New York, NY 10020

First Simon & Schuster hardcover edition July 2024

SIMON & SCHUSTER and colophon are
registered trademarks of Simon & Schuster, LLC

Simon & Schuster: Celebrating 100 Years of Publishing in 2024

For information about special discounts for bulk purchases,
please contact Simon & Schuster Special Sales
at 1-866-506-1949 or business@simonandschuster.com.

The Simon & Schuster Speakers Bureau can bring authors to your live event.
For more information or to book an event,
contact the Simon & Schuster Speakers Bureau
at 1-866-248-3049 or visit our website at www.simonspeakers.com.

Interior design by Ruth Lee-Mui

Manufactured in the United States of America

1 3 5 7 9 10 8 6 4 2

Library of Congress Cataloging-in-Publication Data has been applied for.

ISBN 978-1-6680-5332-4
ISBN 978-1-6680-6533-4 (Int Exp)
ISBN 978-1-6680-5334-8 (ebook)

For Victoria, Cordelia, and Gabriel

CONTENTS

INTRODUCTION: AI'S LIGHT BULB MOMENT

They are called *light bulb moments*: the instant an idea materializes from the dark corners of the unconscious and assumes tangible, almost physical form. Individuals experience light bulb moments, and so do whole societies, moments of public realization that mark the borderline between one era and the next. Such moments occur when a technology is married to a device or software that people without any specialized knowledge can use, and which instantly gives them a new view of the future.

The light bulb itself sparked one such moment. On December 31, 1879, Thomas Edison flicked a switch in his lab in Menlo Park, New Jersey, and his incandescent light bulb illuminated a brave new world: one powered by electricity, not gas or steam. Inventors and scientists on both sides of the Atlantic had been racing one another for close to a decade to commercialize electric power. But many of their inventions were industrial, technical, esoteric. The light bulb was different. Here was a device everyone could hold and behold—and understand.

We've experienced a succession of light bulb moments around digital technology in the past three decades: the internet existed before the web browser, but it was only after the launch of Netscape Navigator in 1994 that the internet era truly dawned. There were MP3 players before the iPod debuted in 2001, but they didn't spark the digital music revolution. There were smartphones before Apple dropped the iPhone in 2007—but

before the iPhone, there wasn't an app for that. On November 30, 2022, artificial intelligence had its light bulb moment. That's the day OpenAI debuted ChatGPT.

As a journalist, I have been covering artificial intelligence since 2016 and have watched as the technology made continual strides. AI has long been a topic of fascination among business executives and technologists, but before ChatGPT, it never quite made it to the center ring of public conversation. For those of us closely following the technology, progress seemed incremental. When ChatGPT debuted, I thought it would be yet another shuffle forward. After all, it didn't seem much different from an AI model called Instruct GPT that OpenAI had launched almost a year before. I had written about Instruct GPT. It was an important refinement of OpenAI's large language model GPT-3. Instruct GPT was easier to control through text-based instructions, or prompts, than GPT-3. It was less likely to produce racist, sexist, homophobic, or otherwise toxic content. And it could do many natural language tasks, from translation to summarization to coding. But Instruct GPT made not the barest ripple in the ocean of public discourse.

So when ChatGPT debuted, I was surprised by the surprise. I shouldn't have been. Form factors—or those interfaces, like Google's search bar, that we use to interact with a technology—matter. The Ford Model T was more than just an internal combustion engine and a set of wheels. And with its simple chatbot interface, OpenAI hit upon a winner. This was a light bulb moment for me too.

During most of the previous decade, AI's advance occurred in narrow domains. A number of the AI achievements I covered before ChatGPT involved software that could beat humans at games. Google's cutting-edge AI lab DeepMind created software called AlphaGo that triumphed over the world champion Lee Sedol at the ancient strategy game Go in March 2016. It was a seminal achievement in computer science: Go has so many possible move combinations that an algorithm cannot use brute calculation to analyze every one, as AI software can with chess. Instead, it has to act on probabilities, learned from games it has played,

to decide which move is best. But games are just games. While they can be proxies for skills we would like an AI system to master, by their very nature games abstract away much of life's complexity. On a Go board, everything is literally black and white, and both players can see the position of every stone.

DeepMind later showcased an AI system that could beat the best human players at a complicated video game called *StarCraft II*, whose imaginary world bears a bit more resemblance to our own messy reality. OpenAI, the same AI research company that would later create ChatGPT, built a team of AI-powered agents that could beat a team of human players at a fast-paced video game called *Dota 2*. But these were not light bulb moments. Too few people were familiar with *StarCraft II* or *Dota 2*. And the software that mastered these games could not be immediately applied to real life.

Before ChatGPT, AI tended to work its magic behind the curtain. It helped familiar products do stuff, like recommend a movie or tag a person in a photo on social media. These AI models were impressive, but they were not general-purpose technologies, and most people didn't consider the products they made possible AI. My editors would sometimes shrug when I'd suggest a story on one of these narrow AI advances, and, in retrospect, I don't blame them. This kind of AI seemed a far cry from the artificial intelligence we'd all seen in movies, the malevolent sentience of HAL 9000 in Stanley Kubrick's *2001: A Space Odyssey*, the benign *Star Trek* computer, or even the beguiling digital assistant in 2013's *Her.*

That started to change in 2018, when AI researchers began making rapid progress with large language models, the type of AI technology that ChatGPT is based on. Their advances showed that a sufficiently large AI system, trained only to predict words in a sentence, could be used for a variety of tasks, from translation to summarization to question-answering. Before, each of those capabilities would have required a different AI system. In the world of AI, this was a big deal, and I wrote about the potentially far-reaching impact on business for *Bloomberg* and then *Fortune.*

But even then, the technology was in the backend of existing software. Because it wasn't embodied in an easy-to-use, consumer-facing product, it was hard for most people—including many experts—to grasp just how much the world was about to change. This tendency to discount progress in AI has a name: the Tesler effect, coined for the computer scientist Larry Tesler, who had quipped that "AI is whatever hasn't been done yet." What has been done, he said, we simply took for granted. What's more, once an AI program could do it, we excised that skill from our definition of "intelligence"—an amorphous quality that required a humanlike combination of abilities that was forever beyond software's grasp.

With ChatGPT, AI finally surmounted the Tesler effect—and then some. Suddenly, people could grasp that AI had arrived. You could talk to it. It could answer, cogently and confidently, if not always correctly. You could ask it for recipes based on the ingredients in your fridge. You could get it to summarize your meeting—in prose, or, if you wished, in verse! You could ask for ideas to improve your business or advice on how to ask for a raise. It could build you a website or write you custom software that could wrangle data and analyze it. Combined with other generative AI models, it could produce images of almost anything you could describe.

Now, AI was getting somewhere. This was more like the AI of Kubrick and *Star Trek*. Our attitude toward the technology transformed from blasé to bewildered overnight. What did this moment mean? What future was AI ushering in? Would we even have a future? There were certainly frightening pronouncements, some from the very executives and researchers closest to the technology's bleeding edge: that AI could displace 300 million jobs worldwide; that ChatGPT was like "a pocket nuclear bomb" that had detonated amid an unsuspecting public; that it would destroy education; that it would destroy trust; that it would obliterate democracy; that even, if we weren't careful, AI could mean "lights out for all of us," as OpenAI's co-founder and chief executive Sam Altman famously warned.

This book is my attempt to answer the many questions posed by

ChatGPT. In researching it, I was struck by the extent to which many of our concerns, as well as some of the moral panic greeting the AI revolution, also attended the introduction of earlier technologies, from the printing press to television to the internet. Some fears about AI date back to the dawn of the computer age. Others have arisen more recently, with the introduction of the internet, GPS, social media, and mobile phones. There are important lessons to be drawn from how we, collectively, met the challenges these earlier inventions posed—in the ways we were changed and, perhaps more importantly, in the ways we weren't. AI represents an acceleration, an extension, and in many cases an exacerbation of trends these earlier technologies set in motion. When it comes to social isolation or distrust in media, AI represents a difference of degree, not necessarily of kind.

But AI is different from these earlier innovations in three key ways. It is a more general-purpose technology than most we have encountered previously. This makes it much more akin to the invention of writing or metal smelting or electricity than to the telephone or even the automobile or airplane. It is, as several thinkers on AI have opined, perhaps "the last invention" we need ever create. That's because it holds the promise of helping to create all future technologies we may require. It has the potential to impact almost every aspect of society.

Second, the pace of AI's progress, as well as its adoption, is faster than prior technologies. ChatGPT reached 100 million users within a month of its release, a milestone that took Facebook, the previous fastest-growing consumer software product, almost four and a half years to reach. Spending on the technology is surging, with more than half of Fortune 500 companies planning to implement ChatGPT-like AI in the coming few years.

The speed at which all this is happening matters because it gives us humans less time: less time to adjust to the new reality AI is helping to create; less time to think about AI's implications and how we want to govern the technology; and less time to act before it becomes too late, before harms, both small and large, occur.

Finally, more than any previous technology, AI strikes at the heart of our perception of what makes us unique as a species: our intelligence and our creativity. When the John Henrys of the world eventually lost out to the steam drill, it devalued human labor. But human physical strength was never dominant. We'd long known creatures that could outrun us, outclimb us, outswim us, and overpower us. But nothing could outthink us. Microprocessors and previous generations of software rivaled some of our cognitive abilities: They could run calculations far faster and more accurately than we could. They could follow any rote set of procedures more exactly and speedier, too. But we *wrote* those rote steps. Computers did not challenge the core of our intellect, our ability to reason, to invent new ways of solving problems, and our creative power to express ideas and emotions in novel ways. AI challenges most, if not all, of these things. This is why it is so profound and unsettling, and why we regard it with such fascination and dread. We are, for once, in a position where we might not be the preeminent intelligence on the planet—at least not for much longer.

While this is vertigo-inducing, we have much to look forward to in this AI-defined future. AI can grant us all superpowers, if we make the right design choices and adopt the right policies. It can alleviate the burden of routine tasks, enabling us to work faster and smarter. It will set off an unprecedented productivity boom that will accelerate economic growth. Rather than destroying education, as some teachers and parents fear, AI will allow each of us to have a personal tutor in our pocket. AI will help us push the boundaries of science and medicine, offering us new cures and more personalized medicine. It will help us advance our understanding of chemistry, biology, and even our own history and prehistory. It will aid us in monitoring our planet and safeguarding biodiversity, as well as helping us use renewable power more efficiently. It will give many of us a constant companion, possibly alleviating loneliness and allowing those who struggle socially to improve their interpersonal skills. Used correctly, it could enhance our democracy.

But there are grave dangers here too. If we make the wrong choices,

AI will enfeeble us individually and collectively. Our social skills may atrophy. At work, we may wind up being disempowered and demotivated, slaves to the machine, rather than its master. Without the right government action, inequality will become worse, and power more concentrated. The explosion of synthetic content may obliterate trust and shred the already threadbare fabric of our democracy. AI's power consumption could hurt our efforts to fight climate change. The technology might help engineer bioweapons. It could make our wars more deadly for civilians and soldiers alike. In the wrong hands, it could be a tool for terror. If we were to automate decisions concerning nuclear weapons, the risk could be world-ending. A future AI superintelligence could even pose a threat to our species.

In the following pages, you will also encounter arguments about AI's impact that don't get as much attention as the stories about the end of democracy or even the world. Chief among them is the threat AI poses to our minds and the way we think. We should not let AI's cognitive prowess diminish our own. Another critical risk that is too little discussed is AI's potential to shove empathy out of the central position it ought to occupy in our decision-making. Finally, AI could worsen the divisions along racial and class lines that already trouble our society.

On a more hopeful note, I also showcase opportunities that are frequently overlooked. Just as AI has the potential to worsen inequality, it will offer us a chance to pull more people into the middle class; to help people create new businesses and become entrepreneurs. It could lead to a flourishing of creativity and culture. And it could improve learning and expand both our personal and collective knowledge.

This exploration of AI begins with a brief history, explaining how we have arrived at the threshold of this new age. We will then examine AI's impacts, starting with how it could profoundly alter our own thinking. Expanding outward, we will explore AI's effects on our social relationships, our jobs, the companies we work for, and the economy as a whole. From there, we will investigate AI's reshaping of art and culture and science and medicine. We will confront AI's complicated role in our efforts

to foster sustainability, the dangers facing democracy, and the threat from the unrestricted use of AI in weapons systems. Finally, we will look at the possibility that AI might result in our extinction.

Across all these dimensions, AI poses unsettling questions around authenticity and trust, and around what philosophers call *ontology* (our beliefs about being and reality) and *epistemology* (how we construct those beliefs). In our AI-driven future, we should insist on maintaining a distinction between the authentic and inauthentic—between a simulation of something and the thing itself. AI is excellent at mimicry, at imitating understanding, writing, thought, and creativity. In most cases, it is incapable of the thing itself—and probably always will be. It can complement and extend key aspects of our humanity, but we should not be fooled into thinking it can substitute for us. We must remember that AI's superpower is its ability to enhance, not replace, our own natural gifts.

In this book, I hope to illuminate a path toward the superpowered future AI can enable. But it is a narrow path. We must step carefully. We must be deliberate in how we build AI models and, as importantly, how we interact with them.

In the end, how we use AI is up to us. We have choices to make as a society—decisions that are rarely discussed in the public realm, whether it's in political campaigns or on earnings calls. Only such careful choices will ensure AI complements our best human skills, rather than supplants them. Government and corporate policies will make all the difference to AI's impact. I outline some of these policies throughout this book.

More than the capabilities of AI models themselves, the precise nature of how we interact with this software will matter tremendously. To get this right, we need to value the insights of researchers who specialize in human–computer interaction and human cognitive biases as much as we do those building ever more powerful AI.

Humans must be teammates with AI software, not mere inspectors of its output. We must insist that processes, not just results, matter. And those processes must encompass empathy—one person's understanding for another, born of lived experience. As humans, this is our most

precious and enduring gift. It is the bedrock of our morality. And it is a thing AI is unlikely to ever possess. We must not let our rush to embrace AI's technical efficiency dislodge empathy, that most human trait, from its preeminence in our civilization. Keeping this lodestar in sight won't be easy. But if we can, we will succeed in mastering AI.

CHAPTER

1

THE CONJURERS

In the spring of 2020, in a vast windowless building the size of twenty football fields on the plains outside Des Moines, Iowa, one of the largest supercomputers ever built crackled to life. The supercomputer was made up of rack upon rack of specialized computer chips—a kind originally developed to handle the intensive work of rendering graphics in video games. More than ten thousand of the chips were yoked together, connected with high-speed fiber-optic cabling. The data center belonged to Microsoft, and it cost hundreds of millions of dollars to construct. But the supercomputer was designed to serve the needs of a small start-up from San Francisco that, at the time, few people outside the niche computer science field known as artificial intelligence had ever heard of. It was called OpenAI.

Microsoft had invested $1 billion in OpenAI in July 2019 to gain access to its technology. As part of the deal, Microsoft promised to build this supercomputer. For the next thirty-four days, the supercomputer worked around the clock training one of the largest pieces of AI software

ever developed. The software could encode the relationship between 175 billion data points. OpenAI fed this AI software text—more than 2.5 billion web pages collected over a decade of web crawling, millions of Reddit conversations, tens of thousands of books, and all of Wikipedia. The software's objective was to create a map of the relationships between all of the words in that vast dataset. This statistical map would allow it to take any arbitrary text sequence and predict the next most likely word. By doing so, it could output many paragraphs—in a variety of genres and styles—that were almost indistinguishable from human writing.

OpenAI called the software GPT-3, and its debut in June 2020 stunned computer scientists. Never before had software been so capable of writing. GPT-3 could do far more than just assemble prose and poetry. It could also code. It could answer factual questions. It was capable of reading comprehension and summarization. It could determine the sentiment of a piece of writing. It could translate between English and French, and French and German, and many other languages. It could answer some questions involving commonsense reasoning. In fact, it could perform dozens of language-related tasks, even though it had only ever been trained to predict the next word in a sequence. It did all this based on instructions that were conversational, the way you would speak to a person. Scientists call this "natural language," to differentiate it from computer code. The potential of GPT-3 convinced Satya Nadella, Microsoft's chief executive, to quietly double, and then triple, his initial $1 billion investment in the small San Francisco AI lab. GPT-3 would, in turn, lead two years later to ChatGPT.

ChatGPT was not the first AI software whose output could pass for human. But it was the first that could be easily accessed by hundreds of millions of people. It awakened the public to AI's possibilities and set off a race among the largest technology companies and well-funded start-ups to turn those possibilities into reality. Within months, Microsoft would commit even more money to OpenAI, another $10 billion, and begin incorporating OpenAI's even more powerful GPT-4 model into products such as Microsoft Word and PowerPoint, which hundreds of

millions of people use daily. Google, scrambling to catch up, created a general-purpose chatbot of its own, Bard, and began integrating generative AI into its search engine, potentially destabilizing the business models of entire industries, from news and media to e-commerce. It also began training even larger, more powerful AI systems that analyze and create images, sounds, and music, not just text. This model, called Gemini, has since replaced Bard and been integrated into several Google products. Meta began releasing powerful AI models freely for anyone to use. Amazon and Apple started building generative AI too.

This competition is propelling us toward a single general-purpose AI system that could equal or exceed human abilities at almost any cognitive task—a moment known as the *singularity*. While many still doubt that moment is at hand, it is closer than ever.

Why did ChatGPT seem so remarkable? Because we could talk to it. That a computer's conversational skills—and not its fluency in multiplying five-digit numbers or detecting patterns in stock market gyrations—should be the yardstick by which we measure machine intelligence has a long and contentious history. The vision of an intelligent digital interlocutor was there at the dawn of the computer age. It has helped guide the entire field of AI—for better, and for worse.

THE TURING TEST—AI'S ORIGINAL SIN

The idea of dialogue as the hallmark of intelligence dates to the mid-twentieth century and the writings of one of that era's most exceptional minds, Alan Turing, the brilliant mathematician best known for helping to crack the Nazis' Enigma code during World War II. In 1936, when he was just twenty-four, Turing proposed the design of a hypothetical device that would serve as the inspiration for modern computers. And it was Turing who, in a 1948 report for a British government laboratory, suggested that computers might one day be considered intelligent. What mattered, he argued, was the quality of a machine's output, not the process that led to it. A machine that could beat a person at chess was more

"intelligent," even if the machine arrived at its moves through a method, brute calculation, that bore little resemblance to the way its human opponent thought.

Two years later, in 1950, Turing expanded on these ideas in his seminal paper entitled "Computer Machinery and Intelligence." He suggested a test for machine intelligence that he called "the Imitation Game." It involved an interrogator asking questions of a person and a computer. All three are isolated from one another in separate rooms. The interrogator receives responses in the form of typewritten notes that are labeled X or Y depending on whether the human or the machine produced them. The interrogator's job is to determine, based on the typed answers, whether X is a person or a machine. Turing proposed that a computer should be considered intelligent if the interrogator can't tell the difference.

Critically, Turing explained that the test was not about the accuracy or truthfulness of answers. He fully expected that to "win" the game, both the human and the machine might lie. Nor was it about specialist knowledge. In describing the game, he postulated that it would be a machine's mastery of *the form* of everyday conversation, as well as an apparent grasp of common sense, that would make it indistinguishable from the human.

The Imitation Game, later dubbed "the Turing Test," has had a profound impact on how computer scientists have regarded machine intelligence. But it proved controversial almost from the start. Wolfe Mays, a philosopher and professor at the University of Manchester, was among the contemporary critics who attacked Turing's test because of its focus on the machine's output, instead of its internal processes. He called machines' cold, logical calculus "the very antithesis of thinking," which Mays argued was a far more mysterious, intuitive phenomenon. Mays was among those who believed intelligence was closely linked to human consciousness, and that consciousness, in turn, could not be reduced to mere physics. Turing, he wrote, seemed "implicitly to assume that the whole of intelligence and thought can be built up summatively from the warp and woof of atomic propositions."

Decades later, the philosopher John Searle constructed a thought experiment to highlight what he saw as the Turing Test's fatal flaw. Searle imagined a man who can neither read nor speak Chinese being locked in a room with a Chinese dictionary, some sheets of paper, and a pencil. Notes with Chinese characters are slipped through a slot in the door, and the man's job is to look up the symbols in the dictionary and transcribe the Chinese characters he finds there onto a fresh slip of paper and push these slips out through the slot. Searle said it would be ridiculous to say that the man in the room "understood" Chinese just because he could produce accurate definitions of Chinese words. So, too, Turing was wrong, Searle wrote, to consider a computer intelligent just because it could mimic the outward characteristics of a human dialogue.

Debates about these issues have become louder with the advent of ultra-large language models, such as GPT-4. It certainly doesn't help that cognitive science has yet to resolve the nature of human intelligence, let alone consciousness. Can intelligence be distilled to a single number—an intelligence quotient, or IQ—based on how someone performs on a standardized test? Or is it more appropriate, as the Harvard psychologist and neuroscientist Howard Gardner argued, to speak of multiple intelligences that take into account the athletic genius of a Michael Jordan, the musical talents of a Taylor Swift, and the interpersonal skills of a Bill Clinton? Neuroscientists and cognitive psychologists continue to argue.

Beyond elevating outcomes over process, the Turing Test's emphasis on deception as the hallmark of intelligence encouraged the use of deceit in testing AI software. This makes the benchmark fundamentally unethical in the eyes of many modern AI ethicists, who blame the Turing Test for encouraging a tradition of AI engineers testing their systems on unwitting humans. In recent years, companies have tested AI software that could play strategy games like Go and Diplomacy against unsuspecting human players. Researchers defended the deception as necessary: humans might alter their playing style and strategies if they knew they were up against a software opponent. Ethicists and journalists lambasted Google when, in 2018, it showed off its new Duplex digital assistant by

having the software phone a restaurant and make a reservation, fooling the maître d' into thinking she was interacting with a person. Today most "responsible AI" policies state that people should always be informed when they're interacting with AI software—and yet companies sometimes find justifications for not doing so.

Perhaps the Turing Test's most enduring and troubling legacy concerns its framing as a game of wits in which a human is pitted against a computer whose goal is to mimic and equal the human. As the technology writer John Markoff notes in his book *Machines of Loving Grace*, one effect of this was that generations of AI researchers would try to create software that matched or exceeded human performance at some tasks, and, as a result, were evaluated for their potential to replace humans. An alternative, Markoff says, would be to view AI software as having different but complementary capabilities. This framework would encourage the development of AI systems designed to augment people, rather than replace them.

But this complementary perspective has struggled to gain traction. When they wish to prove their software's capabilities, AI researchers still tend to follow Turing's example and benchmark their systems against human performance. Some of this is simply good marketing, of course—software besting humans at a familiar board game or some professional exam makes for better headlines. But among AI researchers, the desire to equal or exceed human performance runs deeper than just a clever media strategy, and much of that is down to Turing. More recently, the latest generation of AI language models has been subjected to a battery of tests designed to evaluate people, including the bar exam and the U.S. medical licensing exam. Today's most powerful AI systems have passed all of these tests, often scoring better than the average person. And of course, ChatGPT's mastery of the original Turing Test—its convincing dialogue—has helped make it such a sensation. But dig a little deeper and the software's imitation of us begins to slip. As anyone who has played around with ChatGPT soon discovers, the intellectual abilities of today's generative AI can be frustratingly brittle and weirdly alien: it can

answer a tough question about particle physics with mastery, and then botch a simple logic problem an eight-year-old could ace. Such inconsistency, brilliance and stupidity commingled, is certainly not what Turing had in mind.

BUILDING TURING'S INTELLIGENT MACHINES

The possibility of such a paradox—that a computer could master conversation without any real understanding of what it was saying—wasn't even a distant flicker in the imaginations of the scientists who first took up the challenge of trying to turn Turing's vision of a thinking machine into reality. Around the time Turing dreamed up the Imitation Game, the first universal computers, which owed much to his prewar conceptual designs, were being plugged in. These were colossal machines: ENIAC, one of these early computers, took almost the entire basement of the University of Pennsylvania engineering department, occupying 1,500 square feet and consisting of thirty tons of wiring and vacuum tubes. It could perform five thousand addition calculations per second, which was much faster than previous mechanical calculating machines, but almost one trillion times slower than the powerful chips that run today's AI software. Yet a small group of pioneers believed they could teach machines like ENIAC to think.

John McCarthy, a young mathematics professor at Dartmouth College, was one of them. By the early 1950s, the idea of imbuing these new computers with intelligence began to take hold. As a key step toward that goal, scientists were also trying to discover algorithms that could accurately describe how the human brain itself worked. McCarthy was at the heart of this intellectual fervor, working on what was called *automata theory*. Others were carrying out similar research but calling it cybernetics and information processing. In early 1955, McCarthy had the idea of bringing some cohesion to these efforts by hosting a two-month workshop the following summer at Dartmouth. He would invite about a dozen academics—mostly mathematicians, but also electrical

engineers and psychologists—to figure out how to get a computer to learn and think.

In writing a funding proposal for the workshop, McCarthy coined the term "artificial intelligence" to refer to the emerging field. McCarthy and his fellow workshop organizers set an ambitious agenda. "An attempt will be made to find how to make machines use language, form abstractions and concepts, solve [the] kinds of problems now reserved for humans, and improve themselves," they wrote in their grant application to the Rockefeller Foundation. They believed that, with a "carefully selected" group of scientists working together, "a significant advance could be made in one or more of these problems" that summer.

As it turned out, McCarthy and his fellow organizers had been wildly optimistic. The 1956 workshop failed to unify the field around a singular vision and research program, disappointing McCarthy. Some of the attendees even disliked the name "artificial intelligence," believing it connoted something phony. Heated debate over the term diverted time and attention from more substantive discussions. McCarthy defended his choice—arguing that it emphasized that the field's goal was intelligence, and not merely automation. The name stuck. And, while the Dartmouth workshop failed to braid the various strands of AI research into a single rope, it showcased many of the conceptual threads—and many of the challenges—that would define the pursuit of AI over the next six decades, including one idea that has sparked the current AI revolution.

That idea was called the artificial neural network. By the late 1940s, neuroscientists had discovered that human neurons began life undifferentiated from one another—any difference in the way the mature neurons functioned in the brain must, they concluded, be learned. They also knew that neurons fed electro-chemical signals to other neurons in what seemed to be a sort of hierarchy. These two insights prompted scientists to search for ways to duplicate this structure, first using vacuum tubes and later using digital computers and software.

The earliest neural networks had just two layers. In the input layer, neurons would each take in data from a different portion of whatever

image, audio recording, or text the network was analyzing. Every input neuron applied a mathematical formula to its bit of data and passed the result along to a single output neuron. The output neuron summed the answers from all the input neurons and then applied its own mathematical formula, with the result determining the answer it gave to a user. If that answer was wrong, variables, called weights and biases, that are part of each neuron's formula, were adjusted. The neurons closest to having the right answer had their weights and biases increased, while the others had theirs decreased. Then the whole network tried again with another example from a dataset. This process enabled the network to gradually learn, over the course of many examples, how to produce the right answer.

A number of computer scientists thought these early neural networks would be the way to achieve artificial intelligence. The approach worked well for teaching computers to make simple binary classifications, determining whether something was light or dark, a circle or a square. But it couldn't do much else. Marvin Minsky, a brilliant young Harvard mathematician who helped McCarthy organize the workshop, built one of the first analog neural networks using parts salvaged from a B-24 bomber. He presented his work on neural networks at Dartmouth that summer. But much of Minsky's presentation dwelled on neural networks' many limitations.

Soon Minsky, who went on to found MIT's AI Lab, emerged, with the zeal of a convert, as the technique's harshest critic. He would turn instead to a different idea about how to build machines that could learn and think, one that used logical rules to tell a computer how to reason. For example, to create a system that could identify cars, motorcycles, and bicycles in images, you would program the computer to recognize features that humans knew were important in making this determination—things like handlebars, the number of wheels, doors and windows and exhaust pipes. Then you would write code that would instruct the computer in how to reason about these features: If it has handlebars, but no doors or exhaust pipe, it is a bike. If it has an exhaust pipe, but no doors,

it's a motorcycle. In the years following the Dartmouth workshop, these symbolic reasoning methods yielded what seemed like steady progress toward humanlike AI. Scientists created software that could play backgammon, checkers, and then chess about as well as the average person. Then, in 1966, an AI system using Minsky's rules-based approach came within a whisker of passing the Turing Test.

THE ELIZA EFFECT

The achievement was the work of an engineer working at Minsky's MIT lab named Joseph Weizenbaum. Born in Berlin in 1923 to a well-to-do Jewish family, he wound up in Detroit at thirteen after fleeing the Nazis. Weizenbaum's relationship with his parents was strained and, finding himself socially isolated as he struggled to learn English, he sought refuge first in mathematics, and later on his psychoanalyst's couch. He gravitated to the nascent field of computing and eventually landed a job as a programmer, writing software for the giant computers that General Electric was building in the emerging tech hub of Silicon Valley. He became intrigued by the possibility of interacting with a machine, not in code, but in natural language. Knowing his interests, a mutual friend introduced him to the Stanford University psychiatrist Kenneth Colby, who thought computers might open up new avenues for therapy, as well as providing interesting ways to research mental illness by modeling the human brain. They collaborated on a project to create a kind of software therapist. Weizenbaum took the idea with him when he joined the MIT faculty in 1963. The result, three years later, was the world's first chatbot, Eliza, named after Eliza Doolittle from *Pygmalion*.

Eliza was designed to mimic, in text, the interaction between a patient and a psychoanalyst. Weizenbaum chose this "persona" for Eliza deliberately because it helped mask the chatbot's relatively weak language understanding. The system analyzed the text a user gave it and, according to a complex series of rules, tried to match the text to one of several pre-scripted replies. If it was uncertain, it would sometimes

simply parrot the user's input back in the form of a question, as a psychoanalyst might. In other cases, it would lean on something vague and sympathetic, such as, "I see," or "Go on," or "That's very interesting."

The coding behind Eliza was not particularly sophisticated, even by the standards of the day. Instead, what made Eliza groundbreaking was people's reaction to it. Many of the students whom Weizenbaum observed interacting with Eliza were convinced it was a human therapist. "Some subjects have been very hard to convince that Eliza (with its present script) is *not* human," Weizenbaum wrote. He would later recall that his secretary requested time with Eliza and then, after a bit, requested he leave the room. "I believe this anecdote testifies to the success with which the program maintains the illusion of understanding," he noted. Even computer scientists found themselves making intimate confessions to the chatbot.

In other words, Eliza came pretty close to pulling off the sort of deception Turing envisioned. But Eliza also made glaringly apparent the Turing Test's weaknesses. Chief among these is our own credulity. The people who interacted with Eliza wanted to believe they were talking to a person. This tendency, to anthropomorphize chatbots—and to suspend our own disbelief despite their obvious shortcomings—became known as "the Eliza effect." It remains a powerful force in AI today.

The Eliza effect's biggest impact may have been on Eliza's creator. Rather than experiencing triumph, Weizenbaum was depressed. What did it say about his fellow humans that we were so easily duped? He began to see how difficult it was to distinguish real human thought from its imitation—and yet how vital. Weizenbaum's early collaborator Colby would maintain that if patients found a chatbot like Eliza helpful, then it was a worthwhile invention. But the deception at Eliza's core troubled Weizenbaum. This debate continues today with a new wave of chatbots designed to sound empathetic or be used in therapy.

Minsky, Weizenbaum's MIT colleague, saw no difference between the reasoning a computer performed and the brain—Minsky famously quipped that humans were nothing more than "meat machines."

Weizenbaum made a radical break from this view. He came to see certain realms as beyond the scope of AI because they depended on lived experience—and ultimately emotions—that a computer could never have. Even if the rules of language could be mathematically formalized, something he had come to doubt, AI software could never actually "understand" language. Drawing on the later philosophical writings of Ludwig Wittgenstein, Weizenbaum argued language was insufficient for conveying meaning even between two people—because none of us shares the same lived experience as anyone else—let alone between a person and a piece of software, which has no lived experience. Plus, so much of communication is nonverbal. Weizenbaum related the story of standing next to his wife, looking over their children as they lay asleep in bed. "We spoke to each other in silence, rehearsing a scene as old as mankind itself," he later wrote. Then he quoted the playwright Eugene Ionesco: "Not everything is unsayable in words, only the living truth."

Weizenbaum argued that this fundamental difference between human and machine—the computer's incapacity for true empathy—meant that even if AI could learn to perform every human task as well as or better than a person, it ought to be banned from doing certain things. He thought software should never be allowed to make decisions with irreversible consequences or "where a computer system is proposed as a substitute for a human function that involves interpersonal respect, understanding, and love." He worried that even as we all too readily anthropomorphized our machines, we all too eagerly embraced machines as a metaphor for ourselves, devaluing our humanity.

In 1976, Weizenbaum laid out these critiques in a book, *Computer Power and Human Reason: From Judgment to Calculation*. It was a searing indictment of his own field. He accused his fellow AI researchers of having a "hacker mentality" that was more concerned with getting software to perform some task, rather than trying to answer scientifically meaningful questions about the nature of intelligence and thought. Not only did these so-called computer scientists lack scientific rigor, in his view, but both their approach and the potential impact of the AI software

they were trying to build were antihuman. Weizenbaum, who had a commitment to left-wing causes and civil rights, was particularly disturbed by the U.S. government's use of early AI software in the Vietnam War.

Weizenbaum recognized that many AI systems were so complex that their decision-making was inscrutable, even to their creators. He worried that made these systems difficult to control. But he also recognized that people could use the software's opacity as a convenient excuse to duck accountability. The political economy of AI alarmed him, too. In his neo-Marxist critique, postwar American capitalism might have collapsed under its own weight if the computer had not saved it. Software allowed big business and the federal government to exercise control in ways that would have been impossible before. "The computer, then, was used to conserve America's social and political institutions," he wrote. "It buttressed them and immunized them, at least temporarily, against enormous pressures for change." AI made totalitarian government easier, a concern heightened by Weizenbaum's own experience as a refugee from Nazi Germany.

Weizenbaum's fellow computer scientists dismissed *Computer Power and Human Reason* as anti-scientific and vitriolic. But most of his criticisms of AI, and of those building it, remain valid—in fact, his nightmares about AI are more germane than ever.

RESURRECTING THE NEURAL NETWORK

The reason it was easy to ignore Weizenbaum's critiques in the 1970s and it is impossible to do so now has everything to do with one old idea, the neural network, and three new developments: vastly more powerful computer chips, oceans of digital data made easily available thanks to the internet, and a dash of algorithmic innovation.

Most computer scientists had abandoned work on neural networks by the early 1970s, in no small measure due to Minsky's harsh criticism. But a small group of iconoclasts continued to pursue the idea. In the mid-1980s, one of them, a psychologist named David Rumelhart at

the University of California, San Diego, alongside his colleague Ronald Williams and a young English postdoctoral student named Geoffrey Hinton, hit upon a breakthrough that finally allowed neural networks to start doing some interesting things. Their breakthrough was called backpropagation—or backprop, for short.

Backprop solved a key problem. One reason the earliest neural networks could only learn simple binary classifications was because they had just two layers. It turned out that adding additional layers of neurons between the input neurons and the output neuron enabled the network to make much more complex determinations. By the 1980s, the neural networks that Rumelhart, Hinton, and Williams were experimenting with had several layers. But the middle layers of neurons created a problem. In a simple two-layer neural network, figuring out how to adjust the settings—the weights and biases—of each input neuron so the network could learn during training was fairly simple. But with the addition of these extra layers, knowing how to credit each neuron for its contribution to the output became a daunting challenge. Backprop used ideas from calculus to solve this problem. It allowed the weights and biases of every neuron to be adjusted in a way that would allow a multilayered network to learn effectively.

Thanks to backprop, multilayer neural networks could decipher handwriting on envelopes and checks, learn the relationships between people in a family tree, power a program that could recognize printed words and read them aloud through a voice synthesizer, and even keep an early self-driving car between the lane lines on a highway. But there was still a lot that neural networks could not do. They couldn't classify complex objects in images—no cat or dog detector yet. They struggled with speech recognition and language translation. And there were other drawbacks: training neural networks required vast amounts of data, and in many domains enough data simply wasn't available. What's more, processing all that data in large neural networks was agonizingly slow on the computer chips of the day. At the same time, other machine-learning methods using advanced statistics were showing promise. The result was

that many AI researchers and engineers once again wrote off neural networks as a dead end.

DEEP LEARNING TAKES OFF

Hinton, who had now moved to the University of Toronto, remained convinced that neural networks were the right path to humanlike AI. He kept plugging away. In 2004, he founded a research collective devoted to continuing work on neural networks, an approach he soon rebranded as "deep learning." It was clever marketing: *deep* referred simply to the multiple layers of a neural network, but it also implied that neural networks somehow arrived at insights that were more profound than other "shallow" machine-learning methods. Deep learning was about to live up to the promise that had eluded neural networks for more than fifty years.

What changed were two things: the internet made vast amounts of information available, solving neural networks' hunger for data; and a new kind of computer chip, called a "graphics processing unit," or GPU, hit the market. The American semiconductor company Nvidia debuted the first GPU, the GeForce 256, in late 1999. It was a separate printed circuit board, or card, that could be inserted into a computer or data server. It was designed to handle the fast graphics rendering necessary for video games. It did this by computing several data streams in parallel, not sequentially, and by handling other key computations directly on the graphics card, rather than handing these off to a more general-purpose central processing unit (or CPU), as had been the case with a previous generation of graphics cards. GPUs drew a lot of power and produced a lot of heat; each card came with its own fan. But the new chips made home gaming consoles like the Xbox and PlayStation possible. And they were about to change AI research forever.

In 2005, a team of researchers at the University of Maryland showed that the parallel processing power of GPUs might work well for a simple two-layer neural network. The following year, a team at Microsoft demonstrated that the newest GPUs could also much more efficiently