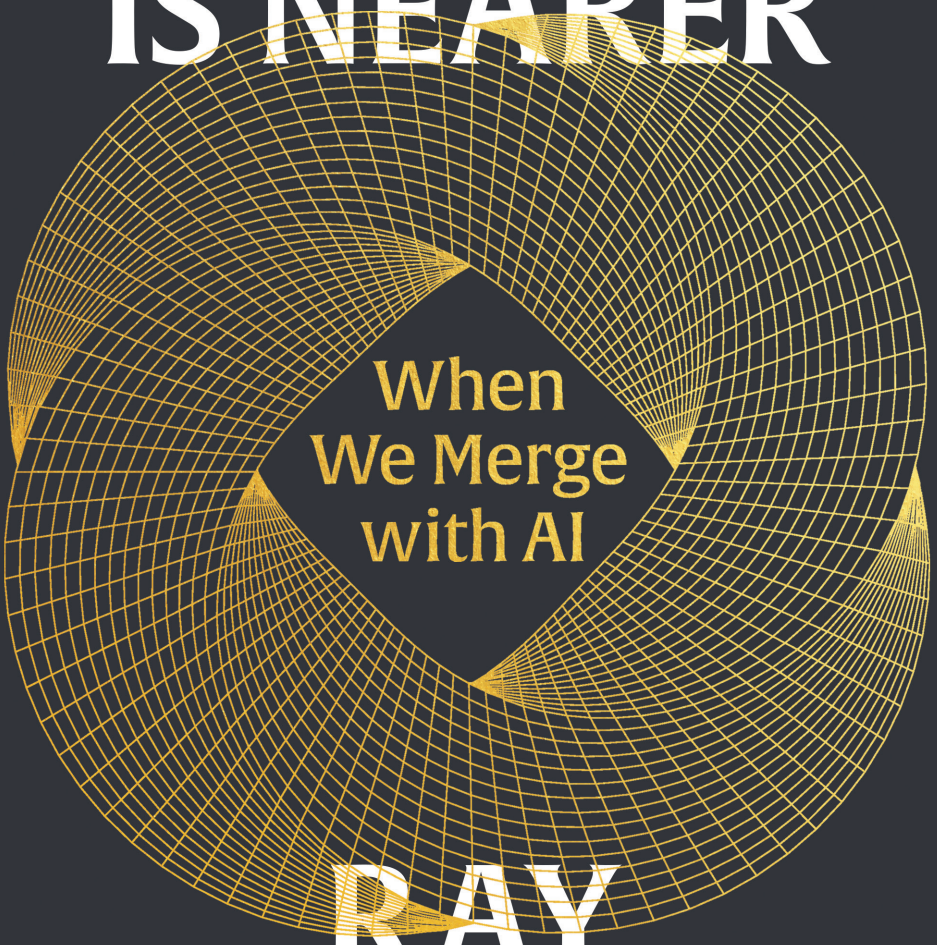
‘A fascinating exploration of our future’
YUVAL NOAH HARARI

THE SINGULARITY IS NEARER



RAY KURZWEIL

‘The best person I know at predicting the future of artificial intelligence’ BILL GATES

THE SINGULARITY IS NEARER

ALSO BY RAY KURZWEIL

The Age of Intelligent Machines

The 10% Solution for a Healthy Life

The Age of Spiritual Machines

Fantastic Voyage (with Terry Grossman, MD)

The Singularity is Near

Transcend (with Terry Grossman, MD)

How to Create a Mind

Danielle: Chronicles of a Superheroine

A Chronicle of Ideas

THE SINGULARITY IS NEARER

WHEN WE MERGE WITH AI

RAY KURZWEIL

THE BODLEY HEAD
LONDON

1 3 5 7 9 10 8 6 4 2

The Bodley Head, an imprint of Vintage, is part of the Penguin Random House group of companies whose addresses can be found at global.penguinrandomhouse.com

Penguin Random House UK

First published in The United States by Viking in 2024
First published in Great Britain by The Bodley Head in 2024

penguin.co.uk/vintage

Printed and bound in Great Britain by Clays Ltd, Elcograf S.p.A.

The authorised representative in the EEA is Penguin Random House Ireland, Morrison Chambers, 32 Nassau Street, Dublin D02 YH68

A CIP catalogue record for this book is available from the British Library

HB ISBN 9781847928290
TPB ISBN 9781847928306

Penguin Random House is committed to a sustainable future for our business, our readers and our planet. This book is made from Forest Stewardship Council® certified paper.

MIX
Paper | Supporting responsible forestry
FSC® C018179
www.fsc.org

To Sonya Rosenwald Kurzweil.
As of a few days ago,
I have now gotten to know her
(and love her) for fifty years!

CONTENTS

ACKNOWLEDGMENTS

I'd like to express my gratitude to my wife, Sonya, for her loving patience through the vicissitudes of the creative process and for sharing ideas with me for fifty years.

To my children, Ethan and Amy; my daughter-in-law, Rebecca; my son-in-law, Jacob; my sister, Enid; and my grandchildren, Leo, Naomi, and Quincy for their love, inspiration, and great ideas.

To my late mother, Hannah, and my late father, Fredric, who taught me the power of ideas in walks through the New York woods, and gave me the freedom to experiment at a young age.

To John-Clark Levin for his meticulous research and intelligent analysis of the data that serves as a basic foundation of this book.

To my longtime editor at Viking, Rick Kot, for his leadership, unwavering guidance, and expert editing.

To Nick Mullendore, my literary agent, for his astute and enthusiastic guidance.

To Aaron Kleiner, my lifelong business partner (since 1973), for his devoted collaboration for the past fifty years.

To Nanda Barker-Hook for her skilled writing assistance and expert oversight and management of my speeches.

To Sarah Black for her outstanding research insights and organization of ideas.

To Celia Black-Brooks for her thoughtful support and expert strategy on sharing my ideas with the world.

To Denise Scutellaro for her adept handling of my business operations.

To Laksman Frank for his excellent graphic design and illustrations.

To Amy Kurzweil and Rebecca Kurzweil for their guidance on the craft of writing, and their own wonderful examples of very successful books.

To Martine Rothblatt for her dedication to all of the technologies I discuss in the book and for our longtime collaborations in developing outstanding examples in these areas.

To the Kurzweil team, who provided significant research, writing, and logistical support for this project, including Amara Angelica, Aaron Kleiner, Bob Beal, Nanda Barker-Hook, Celia Black-Brooks, John-Clark Levin, Denise Scutellaro, Joan Walsh, Marylou Sousa, Lindsay Boffoli, Ken Linde, Laksman Frank, Maria Ellis, Sarah Black, Emily Brangan, and Kathryn Myronuk.

To the dedicated team at Viking Penguin for all of their thoughtful expertise, including Rick Kot, executive editor; Allison Lorentzen, executive editor; Camille LeBlanc, associate editor; Brian Tart, publisher; Kate Stark, associate publisher; Carolyn Coleburn, executive publicist; and Mary Stone, marketing director.

To Peter Jacobs of CAA for his invaluable leadership and support of my speaking engagements.

To the teams at Fortier Public Relations and Book Highlight for their exceptional public relations expertise and strategic guidance in sharing this book far and wide.

To my in-house and lay readers, who have provided many clever and creative ideas.

And, finally, to all the people who have the courage to question outdated assumptions and use their imaginations to do things that have never been done before. You inspire me.

INTRODUCTION

In my 2005 book *The Singularity Is Near*, I set forth my theory that convergent, exponential technological trends are leading to a transition that will be utterly transformative for humanity. There are several key areas of change that are continuing to accelerate simultaneously: computing power is becoming cheaper, human biology is becoming better understood, and engineering is becoming possible at far smaller scales. As artificial intelligence grows in ability and information becomes more accessible, we are integrating these capabilities ever more closely with our natural biological intelligence. Eventually nanotechnology will enable these trends to culminate in directly expanding our brains with layers of virtual neurons in the cloud. In this way we will merge with AI and augment ourselves with millions of times the computational power that our biology gave us. This will expand our intelligence and consciousness so profoundly that it's difficult to comprehend. This event is what I mean by the Singularity.

The term "singularity" is borrowed from mathematics (where it refers to an undefined point in a function, like when dividing by zero) and physics (where it refers to the infinitely dense point at the center of a black hole, where the normal laws of physics break down). But it is important to remember that I use the term as a metaphor. My prediction of the technological Singularity does not suggest that rates of change will actually become infinite, as exponential growth does not imply infinity, nor does a physical singularity. A black hole has gravity strong enough to trap even light itself, but there is no means in quantum

mechanics to account for a truly infinite amount of mass. Rather, I use the singularity metaphor because it captures our inability to comprehend such a radical shift with our current level of intelligence. But as the transition happens, we will enhance our cognition quickly enough to adapt.

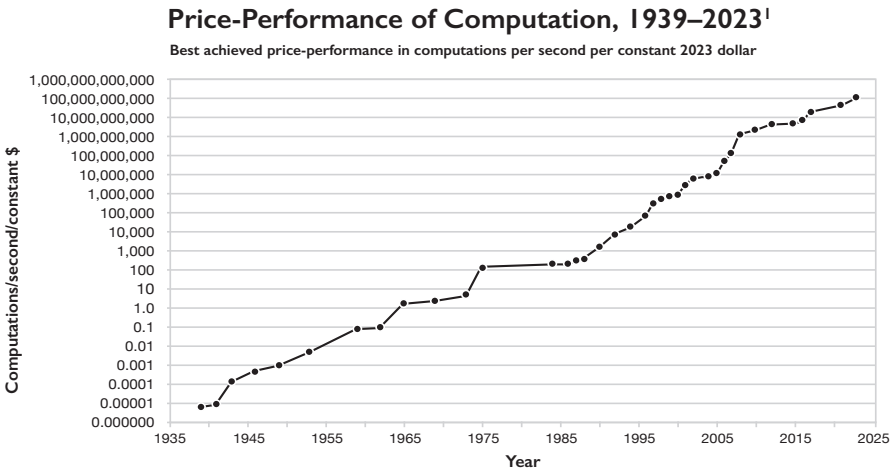
As I detailed in *The Singularity Is Near*, long-term trends suggest that the Singularity will happen around 2045. At the time that book was published, that date lay forty years—two full generations—in the future. At that distance I could make predictions about the broad forces that would bring about this transformation, but for most readers the subject was still relatively far removed from daily reality in 2005. And many critics argued then that my timeline was overoptimistic, or even that the Singularity was impossible.

Since then, though, something remarkable has happened. Progress has continued to accelerate in defiance of the doubters. Social media and smartphones have gone from virtually nonexistent to all-day companions that now connect a majority of the world's population. Algorithmic innovations and the emergence of big data have allowed AI to achieve startling breakthroughs sooner than even experts expected—from mastering games like *Jeopardy!* and Go to driving automobiles, writing essays, passing bar exams, and diagnosing cancer. Now, powerful and flexible large language models like GPT-4 and Gemini can translate natural-language instructions into computer code—dramatically reducing the barrier between humans and machines. By the time you read this, tens of millions of people likely will have experienced these capabilities firsthand. Meanwhile, the cost to sequence a human's genome has fallen by about 99.997 percent, and neural networks have begun unlocking major medical discoveries by simulating biology digitally. We're even gaining the ability to finally connect computers to brains directly.

Underlying all these developments is what I call the law of accelerating returns: information technologies like computing get exponentially cheaper because each advance makes it easier to design the next stage of their own evolution. As a result, as I write this, one dollar buys

about 11,200 times as much computing power, adjusting for inflation, as it did when *The Singularity Is Near* hit shelves.

The following graph, which I'll discuss in depth later in the book, summarizes the most important trend powering our technological civilization: the long-term exponential growth (shown as a roughly straight line on this logarithmic scale) in the amount of computing power a constant dollar can purchase. Moore's law famously observes that transistors have been steadily shrinking, allowing computers to get ever more powerful—but that is just one manifestation of the law of accelerating returns, which already held true long before transistors were invented and can be expected to continue even after transistors reach their physical limits and are succeeded by new technologies. This trend has defined the modern world, and almost all the coming breakthroughs discussed in this book will be enabled by it directly or indirectly.



To maximize comparability of machines, this graph focuses on price-performance during the era of programmable computers, but approximations for earlier electromechanical computing devices show that this trend stretches back at least to the 1880s.[2]

So we have kept on schedule for the Singularity. The urgency of this book comes from the nature of exponential change itself. Trends that were barely noticeable at the start of this century are now actively

impacting billions of lives. In the early 2020s we entered the sharply steepening part of the exponential curve, and the pace of innovation is affecting society like never before. For perspective, the moment you're reading this is probably closer to the creation of the first superhuman AI than to the release of my last book, 2012's *How to Create a Mind*. And you're probably closer to the Singularity than to the release of my 1999 book *The Age of Spiritual Machines*. Or, measured in terms of human life, babies born today will be just graduating college when the Singularity happens. This is, on a very personal level, a different kind of "near" than it was in 2005.

That is why I've written this book now. Humanity's millennia-long march toward the Singularity has become a sprint. In the introduction to *The Singularity Is Near*, I wrote that we were then "in the early stages of this transition." Now we are entering its culmination. That book was about glimpsing a distant horizon—this one is about the last miles along the path to reach it.

Luckily, we can now see this path much more clearly. Although many technological challenges remain before we can achieve the Singularity, its key precursors are rapidly moving from the realm of theoretical science to active research and development. During the coming decade, people will interact with AI that can seem convincingly human, and simple brain–computer interfaces will impact daily life much like smartphones do today. A digital revolution in biotech will cure diseases and meaningfully extend people's healthy lives. At the same time, though, many workers will feel the sting of economic disruption, and all of us will face risks from accidental or deliberate misuse of these new capabilities. During the 2030s, self-improving AI and maturing nanotechnology will unite humans and our machine creations as never before—heightening both the promise and the peril even further. If we can meet the scientific, ethical, social, and political challenges posed by these advances, by 2045 we will transform life on earth profoundly for the better. Yet if we fail, our very survival is in question. And so this book is about our final approach to the Singularity—the opportunities

and dangers we must confront together over the last generation of the world as we knew it.

To begin, we'll explore how the Singularity will actually happen, and put this in the context of our species' long quest to reinvent our own intelligence. Creating sentience with technology raises important philosophical questions, so we'll address how this transition affects our own identity and sense of purpose. Then we will turn to the practical trends that will characterize the coming decades. As I will show, the law of accelerating returns is driving exponential improvements across a very wide range of metrics that reflect human well-being. One of the most obvious downsides of innovation, though, is unemployment caused by automation in its various forms. While these harms are real, we'll see why there is good reason for long-term optimism—and why we are ultimately not in competition with AI.

As these technologies unlock enormous material abundance for our civilization, our focus will shift to overcoming the next barrier to our full flourishing: the frailties of our biology. So next, we'll look ahead to the tools we'll use over the coming decades to gain increasing mastery over biology itself—first by defeating the aging of our bodies and then by augmenting our limited brains and ushering in the Singularity. Yet these breakthroughs may also put us in jeopardy. Revolutionary new systems in biotechnology, nanotechnology, or artificial intelligence could possibly lead to an existential catastrophe like a devastating pandemic or a chain reaction of self-replicating machines. We'll conclude with an assessment of these threats, which warrant careful planning, but as I'll explain, there are very promising approaches for how to mitigate them.

These are the most exciting and momentous years in all of history. We cannot say with confidence what life will be like after the Singularity. But by understanding and anticipating the transitions leading up to it, we can help ensure that humanity's final approach will be safe and successful.

CHAPTER 1

WHERE ARE WE IN THE SIX STAGES?

In *The Singularity Is Near*, I described the basis of consciousness as information. I cited six epochs, or stages, from the beginning of our universe, with each stage creating the next stage from the information processing of the last. Thus, the evolution of intelligence works via an indirect sequence of other processes.

The First Epoch was the birth of the laws of physics and the chemistry they make possible. A few hundred thousand years after the big bang, atoms formed from electrons circling around a core of protons and neutrons. Protons in a nucleus seemingly should not be so close together, because the electromagnetic force tries to drive them violently apart. However, there happens to be a separate force called the strong nuclear force, which keeps the protons together. "Whoever" designed the rules of the universe provided this additional force, otherwise evolution through atoms would have been impossible.

Billions of years later, atoms formed molecules that could represent elaborate information. Carbon was the most useful building block, in that it could form four bonds, as opposed to one, two, or three for many other nuclei. That we live in a world that permits complex chemistry is extremely unlikely. For example, if the strength of gravity were ever so slightly weaker, there would be no supernovas to create the chemical elements that life is made from. If it were just slightly stronger, stars would burn out and die before intelligent life could form. Just this one physical constant had to be in an extremely narrow range

or we would not be here. We live in a universe that is very precisely balanced to allow a level of order that has enabled evolution to unfold.

Several billion years ago, the Second Epoch began: life. Molecules became complex enough to define an entire organism in one molecule. Thus, living creatures, each with their own DNA, were able to evolve and spread.

In the Third Epoch, animals described by DNA then formed brains, which themselves stored and processed information. These brains gave evolutionary advantages, which helped brains develop more complexity over millions of years.

In the Fourth Epoch, animals used their higher-level cognitive ability, along with their thumbs, to translate thoughts into complex actions. This was humans. Our species used these abilities to create technology that was able to store and manipulate information—from papyrus to hard drives. These technologies augmented our brains' abilities to perceive, recall, and evaluate information patterns. This is another source of evolution that itself is far greater than the level of progress before it. With brains, we added roughly one cubic inch of brain matter every 100,000 years, whereas with digital computation we are doubling price-performance about every sixteen months.

In the Fifth Epoch, we will directly merge biological human cognition with the speed and power of our digital technology. This is brain–computer interfaces. Human neural processing happens at a speed of several hundred cycles per second, as compared with several billion per second for digital technology. In addition to speed and memory size, augmenting our brains with nonbiological computers will allow us to add many more layers to our neocortices—unlocking vastly more complex and abstract cognition than we can currently imagine.

The Sixth Epoch is where our intelligence spreads throughout the universe, turning ordinary matter into computronium, which is matter organized at the ultimate density of computation.

In my 1999 book *The Age of Spiritual Machines*, I predicted that a

Turing test—wherein an AI can communicate by text indistinguishably from a human—would be passed by 2029. I repeated that in 2005's *The Singularity Is Near.* Passing a valid Turing test means that an AI has mastered language and commonsense reasoning as possessed by humans. Turing described his concept in 1950,[1] but he did not specify how the test should be administered. In a bet that I have with Mitch Kapor, we defined our own rules that are much more difficult than other interpretations.

My expectation was that in order to pass a valid Turing test by 2029, we would need to be able to attain a great variety of intellectual achievements with AI by 2020. And indeed, since that prediction, AI has mastered many of humanity's toughest intellectual challenges—from games like *Jeopardy!* and Go to serious applications like radiology and drug discovery. As I write this, top AI systems like Gemini and GPT-4 are broadening their abilities to many different domains of performance—encouraging steps on the road to general intelligence.

Ultimately, when a program passes the Turing test, it will actually need to make itself appear far less intelligent in many areas because otherwise it would be clear that it is an AI. For example, if it could correctly solve any math problem instantly, it would fail the test. Thus, at the Turing test level, AIs will have capabilities that in fact go far beyond the best humans in most fields.

Humans are now in the Fourth Epoch, with our technology already producing results that exceed what we can understand for some tasks. For the aspects of the Turing test that AI has not yet mastered, we are making rapid and accelerating progress. Passing the Turing test, which I have been anticipating for 2029, will bring us to the Fifth Epoch.

A key capability in the 2030s will be to connect the upper ranges of our neocortices to the cloud, which will directly extend our thinking. In this way, rather than AI being a competitor, it will become an extension of ourselves. By the time this happens, the nonbiological

portions of our minds will provide thousands of times more cognitive capacity than the biological parts.

As this progresses exponentially, we will extend our minds many millions-fold by 2045. It is this incomprehensible speed and magnitude of transformation that will enable us to borrow the singularity metaphor from physics to describe our future.

CHAPTER 2

REINVENTING INTELLIGENCE

WHAT DOES IT MEAN TO REINVENT INTELLIGENCE?

If the whole story of the universe is one of evolving paradigms of information processing, the story of humanity picks up more than halfway through. Our chapter in this larger tale is ultimately about our transition from animals with biological brains to transcendent beings whose thoughts and identities are no longer shackled to what genetics provides. In the 2020s we are about to enter the last phase of this transformation—reinventing the intelligence that nature gave us on a more powerful digital substrate, and then merging with it. In so doing, the Fourth Epoch of the universe will give birth to the Fifth.

But how will this happen more concretely? To understand what reinventing intelligence entails, we will first look back to the birth of AI and the two broad schools of thought that emerged from it. To see why one prevailed over the other, we will relate this to what neuroscience tells us about how the cerebellum and the neocortex gave rise to human intelligence. After surveying how deep learning is currently re-creating the powers of the neocortex, we can assess what AI still needs to achieve to reach human levels, and how we will know when it has. Finally, we'll turn to how, aided by superhuman AI, we will engineer brain–computer interfaces that vastly expand our neocortices with layers of virtual neurons. This will unlock entirely new modes of

thought and ultimately expand our intelligence millions-fold: this is the Singularity.

THE BIRTH OF AI

In 1950, the British mathematician Alan Turing (1912–1954) published an article in *Mind* titled "Computing Machinery and Intelligence."[1] In it, Turing asked one of the most profound questions in the history of science: "Can machines think?" While the idea of thinking machines dates back at least as far as the bronze automaton Talos in Greek myth,[2] Turing's breakthrough was boiling the concept down to something empirically testable. He proposed using the "imitation game"—which we now know as the Turing test—to determine whether a machine's computation was able to perform the same cognitive tasks that our brains can. In this test, human judges interview both the AI and human foils using instant messaging without seeing whom they are talking to. The judges then pose questions about any subject matter or situation they wish. If after a certain period of time the judges are unable to tell which was the AI responder and which were the humans, then the AI is said to have passed the test.

By transforming this philosophical idea into a scientific one, Turing generated tremendous enthusiasm among researchers. In 1956, mathematics professor John McCarthy (1927–2011) proposed a two-month, ten-person study to be conducted at Dartmouth College, in Hanover, New Hampshire.[3] The goal was the following:

> The study is to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it. An attempt will be made to find how to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves.[4]

In preparing for the conference, McCarthy proposed that this field, which would ultimately automate every other field, be called "artificial intelligence."[5] This is not a designation I like, given that "artificial" makes this form of intelligence seem "not real," but it is the term that has endured.

The study was conducted, but its goal—specifically, getting machines to understand problems described in natural language—was not achieved within the two-month time frame. We are still working on it—of course, now with far more than ten people. According to Chinese tech giant Tencent, in 2017 there were already about 300,000 "AI researchers and practitioners" worldwide,[6] and the 2019 *Global AI Talent Report*, by Jean-François Gagné, Grace Kiser, and Yoan Mantha, counted some 22,400 AI experts publishing original research—of whom around 4,000 were judged to be highly influential.[7] And according to Stanford's Institute for Human-Centered Artificial Intelligence, AI researchers in 2021 generated more than 496,000 publications and over 141,000 patent filings.[8] In 2022, global corporate investment in AI was $189 billion, a thirteenfold increase over the past decade.[9] The numbers will be even higher by the time you read this.

All this would have been hard to imagine in 1956. Yet the Dartmouth workshop's goal was roughly equivalent to creating an AI that could pass the Turing test. My prediction that we'll achieve this by 2029 has been consistent since my 1999 book *The Age of Spiritual Machines*, published at a time when many observers thought this milestone would *never* be reached.[10] Until recently this projection was considered extremely optimistic in the field. For example, a 2018 survey found an aggregate prediction among AI experts that human-level machine intelligence would not arrive until around 2060.[11] But the latest advances in large language models have rapidly shifted expectations. As I was writing early drafts of this book, the consensus on Metaculus, the world's top forecasting website, hovered between the 2040s and the 2050s. But surprising AI progress over the past two years upended expectations, and by May 2022 the Metaculus consensus

exactly agreed with me on the 2029 date.[12] Since then it has even fluctuated to as soon as 2026, putting me technically in the slow-timelines camp![13]

Even experts in the field have been surprised by many of the recent breakthroughs in AI. It's not just that they are happening sooner than most expected, but that they seem to occur suddenly, and without much warning that a leap forward is imminent. For example, in October 2014 Tomaso Poggio, an MIT expert on AI and cognitive science, said, "The ability to describe the content of an image would be one of the most intellectually challenging things of all for a machine to do. We will need another cycle of basic research to solve this kind of question."[14] Poggio estimated that this breakthrough was at least two decades away. The very next month, Google debuted object recognition AI that could do just that. When *The New Yorker*'s Raffi Khatchadourian asked him about this, Poggio retreated to a more philosophical skepticism about whether this ability represented true intelligence. I point this out not as a criticism of Poggio but rather as an observation of a tendency we all share. Namely, before AI achieves some goal, that goal seems extremely difficult and singularly human. But after AI reaches it, the accomplishment diminishes in our human eyes. In other words, our true progress is actually more significant than it seems in hindsight. This is one reason why I remain optimistic about my 2029 prediction.

So why have these sudden breakthroughs occurred? The answer lies in a theoretical problem dating back to the dawn of the field. In 1964, when I was in high school, I met with two artificial intelligence pioneers: Marvin Minsky (1927–2016), who co-organized the Dartmouth College workshop on AI, and Frank Rosenblatt (1928–1971). In 1965 I enrolled at MIT and began studying with Minsky, who was doing foundational work that underlies the dramatic AI breakthroughs we are seeing today. Minsky taught me that there are two techniques for creating automated solutions to problems: the symbolic approach and the connectionist approach.

The symbolic approach describes in rule-based terms how a human

expert would solve a problem. In some cases the systems based on it could be successful. For example, in 1959 the RAND Corporation introduced the "General Problem Solver" (GPS)—a computer program that could combine simple mathematical axioms to solve logic problems.[15] Herbert A. Simon, J. C. Shaw, and Allen Newell developed the General Problem Solver to have the theoretical ability to solve *any* problem that could be expressed as a set of well-formed formulas (WFFs). In order for the GPS to work, it would have to use one WFF (essentially an axiom) at each stage in the process, methodically building them into a mathematical proof of the answer.

Even if you don't have experience with formal logic or proof-based math, this idea is basically the same as what happens in algebra. If you know that $2 + 7 = 9$, and that an unknown number x added to 7 is 10, you can prove that $x = 3$. But this kind of logic has much broader applications than just solving equations. It's also what we use (without even thinking about it) when we ask ourselves whether something meets a certain definition. If you know that a prime number cannot have any factors other than 1 and itself, and you know that 11 is a factor of 22, and that 1 does not equal 11, you can conclude that 22 is not a prime number. By starting with the most basic and fundamental axioms possible, the GPS could do this sort of calculation for much more difficult questions. Ultimately, this is what human mathematicians do as well—the difference is that a machine can (in theory at least) search through every possible way of combining the fundamental axioms in search of the truth.

To illustrate, if there were ten such axioms available to choose from at each point, and let's say twenty axioms were needed to reach a solution, that would mean there were 10^{20}, or 100 billion billion, possible solutions. We can deal with such big numbers today with modern computers, but this was way beyond what 1959 computational speeds could achieve. That year, the DEC PDP-1 computer could carry out about 100,000 operations per second.[16] By 2023 a Google Cloud A3 virtual machine could carry out roughly 26,000,000,000,000,000,000 operations per second.[17] One dollar now buys around 1.6 *trillion* times

as much computing power as it did when the GPS was developed.[18] Problems that would take tens of thousands of years with 1959 technology now take only minutes on retail computing hardware. To compensate for its limitations, the GPS had heuristics programmed that would attempt to rank the priority of possible solutions. The heuristics worked some of the time, and their successes supported the idea that a computerized solution could ultimately solve any rigorously defined problem.

Another example was a system called MYCIN, which was developed during the 1970s to diagnose and recommend remedial treatments for infectious diseases. In 1979 a team of expert evaluators compared its performance with that of human doctors and found that MYCIN did as well as or better than any of the physicians.[19]

A typical MYCIN "rule" reads:

IF: 1) The infection that requires therapy is meningitis, and
2) The type of the infection is fungal, and
3) Organisms were not seen on the stain of the culture, and
4) The patient is not a compromised host, and
5) The patient has been to an area that is endemic for coccidiomycoses, and
6) The race of the patient is one of: [B]lack [A]sian [I]ndian, and
7) The cryptococcal antigen in the csf was not positive
THEN: There is suggestive evidence (.5) that cryptococcus is not one of the organisms (other than those seen on cultures or smears) which might be causing the infection.[20]

By the late 1980s these "expert systems" were utilizing probability models and could combine many sources of evidence to make a decision.[21] While a single *if-then* rule would not be sufficient by itself, by combining many thousands of such rules, the overall system could make reliable decisions for a constrained problem.

Although the symbolic approach has been used for over half a

century, its primary limitation has been the "complexity ceiling."[22] When MYCIN and other such systems made a mistake, correcting it might fix that particular issue but would in turn give rise to three other mistakes that would rear their heads in other situations. There seemed to be a limit on intricacy that made the overall range of real-world problems that could be addressed very narrow.

One way of looking at the complexity of rule-based systems is as a set of possible failure points. Mathematically, a group of n things has 2^{n-1} subsets (not counting the empty set). Thus, if an AI uses a rule set with only one rule, there is only one failure point: Does that rule work correctly on its own or not? If there are two rules, there are three failure points: each rule on its own, and interactions in which those two rules are combined. This grows exponentially. Five rules means 31 potential failure points, 10 rules means 1,023, 100 rules means more than one thousand billion billion billion, and 1,000 rules means over a googol googol googols! Thus, the more rules you have already, the more each new rule adds to the number of possible subsets. Even if only an extremely minuscule fraction of possible rule combinations introduce a new problem, there comes a point (where exactly this point lies varies from one situation to another) where adding one new rule to fix a problem is likely to cause more than one additional problem. This is the complexity ceiling.

Probably the longest-running expert system project is Cyc (from the word "encyclopedic"), created by Douglas Lenat and his colleagues at Cycorp.[23] Initiated in 1984, Cyc has the goal of encoding all of "commonsense knowledge"—broadly known facts like *A dropped egg will break* or *A child running through the kitchen with muddy shoes will annoy his parents*. These millions of small ideas are not clearly written down in any one place. They are unspoken assumptions underlying human behavior and reasoning that are necessary for understanding what the average person knows in a variety of domains. Yet because the Cyc system also represents this knowledge with symbolic rules, it, too, has to face the complexity ceiling.

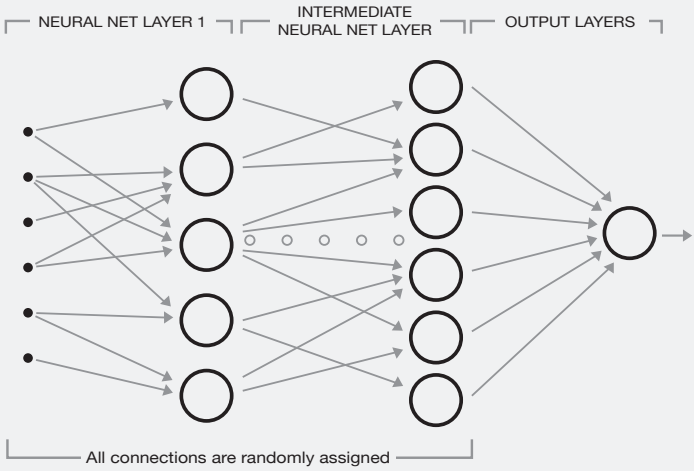
Back in the 1960s, as Minsky advised me on the pros and cons of the

symbolic approach, I began to see the added value of the connectionist one. This entails networks of nodes that create intelligence through their structure rather than through their content. Instead of using smart rules, they use dumb nodes that are arranged in a way that can extract insight from data itself. As a result, they have the potential to discover subtle patterns that would never occur to human programmers trying to devise symbolic rules. One of the key advantages of the connectionist approach is that it allows you to solve problems without understanding them. Even if we had a perfect ability to formulate and implement error-free rules for symbolic AI problem-solving (which we do not), we would be limited by our imperfect understanding of which rules would be optimal in the first place.

This is a powerful way to tackle complex problems, but it is a double-edged sword. Connectionist AI is prone to becoming a "black box"—capable of spitting out the correct answer, but unable to explain how it found it.[24] This has the potential to become a major issue because people will want to be able to see the reasoning behind high-stakes decisions about things like medical treatment, law enforcement, epidemiology, or risk management. This is why many AI experts are now working to develop better forms of "transparency" (or "mechanistic interpretability") in machine learning–based decisions.[25] It remains to be seen how effective transparency will be as deep learning becomes more complex and more powerful.

Back when I started in connectionism, though, the systems were much simpler. The basic idea was to create a computerized model inspired by how human neural networks work. At first this was very abstract because the method was devised before we had a detailed understanding of how biological neural networks are actually organized.

DIAGRAM OF SIMPLE NEURAL NET



Here is the basic schema for a neural net algorithm. Many variations are possible, and the designer of the system needs to provide certain critical parameters and methods (detailed below).

Creating a neural net solution to a problem involves the following steps:

- Define the input.
- Define the topology of the neural net (i.e., the layers of neurons and the connections between the neurons).
- Train the neural net on examples of the problem.
- Run the trained neural net to solve new examples of the problem.
- Take your neural net company public.

These steps (except for the last one) are detailed below:

THE PROBLEM INPUT

The problem input to the neural net consists of a series of numbers. This input can be:

- in a visual pattern recognition system, a two-dimensional array of numbers representing the pixels of an image; or
- in an auditory (e.g., speech) recognition system, a two-dimensional array of numbers representing a sound, in which the first dimension represents parameters of the sound (e.g., frequency components) and the second dimension represents different points in time; or
- in an arbitrary pattern recognition system, an *n*-dimensional array of numbers representing the input pattern.

DEFINING THE TOPOLOGY

To set up the neural net, the architecture of each neuron consists of:

- multiple inputs in which each input is "connected" to either the output of another neuron or one of the input numbers; and
- generally, a single output, which is connected either to the input of another neuron (which is usually in a higher layer) or to the final output.

SET UP THE FIRST LAYER OF NEURONS

- Create N_0 neurons in the first layer. For each of these neurons, "connect" each of the multiple inputs of the neuron to "points" (i.e., numbers) in the problem input. These connections can be determined randomly or using an evolutionary algorithm (see below).
- Assign an initial "synaptic strength" to each connection

created. These weights can start out all the same, can be assigned randomly, or can be determined in another way (see below).

SET UP THE ADDITIONAL LAYERS OF NEURONS

Set up a total of M layers of neurons. For each layer, set up the neurons in that layer.

For $layer_i$:

- Create N_i neurons in $layer_i$. For each of these neurons, "connect" each of the multiple inputs of the neuron to the outputs of the neurons in $layer_{i-1}$ (see variations below).
- Assign an initial "synaptic strength" to each connection created. These weights can start out all the same, can be assigned randomly, or can be determined in another way (see below).
- The outputs of the neurons in $layer_M$ are the outputs of the neural net (see variations below).

THE RECOGNITION TRIALS

HOW EACH NEURON WORKS

Once the neuron is set up, it does the following for each recognition trial:

- Each weighted input to the neuron is computed by multiplying the output of the other neuron (or initial input) that the input to this neuron is connected to by the synaptic strength of that connection.
- All of these weighted inputs to the neuron are summed.

- If this sum is greater than the firing threshold of this neuron, then this neuron is considered to fire and its output is 1. Otherwise, its output is 0 (see variations below).

DO THE FOLLOWING FOR EACH RECOGNITION TRIAL

For each layer, from $layer_0$ to $layer_M$, and
for each neuron in the layer:

- Sum its weighted inputs. (Each weighted input = the output of the other neuron [or initial input] that the input to this neuron is connected to multiplied by the synaptic strength of that connection.)
- If this sum of weighted inputs is greater than the firing threshold for this neuron, set the output of this neuron to 1, otherwise set it to 0.

TO TRAIN THE NEURAL NET

- Run repeated recognition trials on sample problems.
- After each trial, adjust the synaptic strengths of all the interneuronal connections to improve the performance of the neural net on this trial. (See the discussion below on how to do this.)
- Continue this training until the accuracy rate of the neural net is no longer improving (i.e., reaches an asymptote).

KEY DESIGN DECISIONS

In the simple schema above, the designer of this neural net algorithm needs to determine at the outset:

- What the input numbers represent.
- The number of layers of neurons.
- The number of neurons in each layer. (Each layer does not necessarily need to have the same number of neurons.)
- The number of inputs to each neuron in each layer. The number of inputs (i.e., interneuronal connections) can also vary from neuron to neuron and from layer to layer.
- The actual "wiring" (i.e., the connections). For each neuron in each layer, this consists of a list of other neurons, the outputs of which constitute the inputs to this neuron. This represents a key design area. There are a number of possible ways to do this:
  - (i) Wire the neural net randomly; or
  - (ii) Use an evolutionary algorithm (see below) to determine an optimal wiring; or
  - (iii) Use the system designer's best judgment in determining the wiring.
- The initial synaptic strengths (i.e., weights) of each connection. There are a number of possible ways to do this:
  - (i) Set the synaptic strengths to the same value; or
  - (ii) Set the synaptic strengths to different random values; or
  - (iii) Use an evolutionary algorithm to determine an optimal set of initial values; or
  - (iv) Use the system designer's best judgment in determining the initial values.
- The firing threshold of each neuron.
- The output, which can be:
  - (i) the outputs of layer$_M$ of neurons; or
  - (ii) the output of a single output neuron, the inputs of which are the outputs of the neurons in layer$_M$; or

(iii) a function of (e.g., a sum of) the outputs of the neurons in $layer_M$; or

(iv) another function of neuron outputs in multiple layers.

- The synaptic strengths of all the connections, which must be adjusted during the training of this neural net. This is a key design decision and is the subject of a great deal of research and discussion. There are a number of possible ways to do this:

  (i) For each recognition trial, increment or decrement each synaptic strength by a (generally small) fixed amount so that the neural net's output more closely matches the correct answer. One way to do this is to try both incrementing and decrementing and see which has the more desirable effect. This can be time-consuming, so other methods exist for making local decisions on whether to increment or decrement each synaptic strength.

  (ii) Other statistical methods exist for modifying the synaptic strengths after each recognition trial so that the performance of the neural net on that trial more closely matches the correct answer.

  (iii) Note that neural net training will work even if the answers to the training trials are not all correct. This allows using real-world training data that may have an inherent error rate. One key to the success of a neural net–based recognition system is the amount of data used for training. Usually a very substantial amount is needed to obtain satisfactory results. Just as with human students, the amount of time that a neural net spends learning its lessons is a key factor in its performance.

VARIATIONS

Many variations of the above are feasible:

- There are different ways of determining the topology. In particular, the interneuronal wiring can be set either randomly or using an evolutionary algorithm, which mimics the effects of mutation and natural selection on network design.
- There are different ways of setting the initial synaptic strengths.
- The inputs to the neurons in $layer_i$ do not necessarily need to come from the outputs of the neurons in $layer_{i-1}$. Alternatively, the inputs to the neurons in each layer can come from any lower or higher layer.
- There are different ways to determine the final output.
- The method described above results in an "all or nothing" (1 or 0) firing, called a nonlinearity. There are other nonlinear functions that can be used. Commonly, a function is used that goes from 0 to 1 in a rapid but relatively more gradual fashion. Also, the outputs can be numbers other than 0 and 1.
- The different methods for adjusting the synaptic strengths during training represent key design decisions.

The above schema describes a "synchronous" neural net, in which each recognition trial proceeds by computing the outputs of each layer, starting with $layer_0$ through $layer_M$. In a true parallel system, in which each neuron is operating independently of the others, the neurons can operate "asynchronously" (i.e., independently). In an asynchronous approach, each neuron is constantly scanning its inputs and fires whenever the sum of its weighted inputs exceeds its threshold (or whatever its output function specifies).

The goal is to then find actual examples from which the system can figure out how to solve a problem. A typical starting point is to have the neural net wiring and synaptic weights set randomly, so that the answers produced by this untrained neural net will thus also be random. The key function of a neural net is that it must learn its subject matter, just like the mammalian brains on which it is (at least roughly) modeled. A neural net starts out ignorant but is programmed to maximize a "reward" function. It is then fed training data (e.g., photos containing corgis and photos containing no corgis, as labeled by humans in advance). When the neural net produces a correct output (e.g., accurately identifying whether there's a corgi in the image), it gets reward feedback. This feedback can then be used to adjust the strength of each interneuronal connection. Connections that are consistent with the correct answer are made stronger, while those that provide a wrong answer are weakened.

Over time, the neural net organizes itself to be able to provide the correct answers without coaching. Experiments have shown that neural nets can learn their subject matter even with unreliable teachers. If the training data is labeled correctly only 60 percent of the time, a neural net can still learn its lessons with an accuracy well over 90 percent. Under some conditions, even smaller proportions of accurate labels can be used effectively.[26]

It's not intuitive that a teacher can train a student to surpass her own abilities, and likewise it can be confusing how unreliable training data can yield excellent performance. The short answer is that errors can cancel each other out. Let's say you're training a neural net to recognize the numeral 8 from handwritten samples of the numerals 0 through 9. And let's say that a third of the labels are inaccurate—a random mixture of 8s coded as 4s, 5s coded as 8s, and so on. If the dataset is large enough, these inaccuracies will offset each other and not skew the training much in any particular direction. This preserves most of the useful information in the dataset about what 8s look like, and still trains the neural net to a high standard.

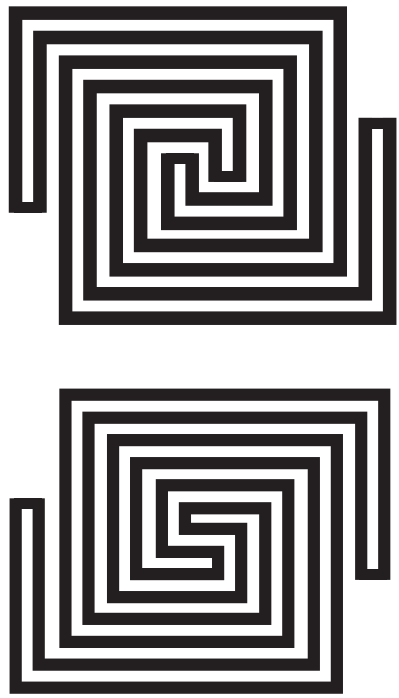
Despite these strengths, early connectionist systems had a funda-

mental limitation. One-layer neural networks were mathematically incapable of solving some kinds of problems.[27] When I visited Professor Frank Rosenblatt at Cornell in 1964, he showed me a one-layer neural network called the Perceptron, which could recognize printed letters. I tried simple modifications to the input. The system did a fairly good job of auto-association (that is, it could recognize the letters even if I covered parts of them) but fared less well with invariance (that is, it failed to recognize letters after size and font changes).

In 1969 Minsky criticized the surge in interest in this area, even though he had done pioneering work on neural nets in 1953. He and Seymour Papert, the two cofounders of the MIT Artificial Intelligence Laboratory, wrote a book called *Perceptrons*, which formally demonstrated why a Perceptron was inherently incapable of determining whether or not a printed image was connected. The two images on page 28 are from the cover of *Perceptrons*. The top image is not connected (the black lines do not form a single contiguous shape), whereas the bottom image is connected (the black lines form a single contiguous shape). A human can determine this, as can a simple software program. A feed-forward (in which connections between the nodes do not form any loops) Perceptron such as Rosenblatt's Mark 1 Perceptron cannot make this determination.

In short, the reason feed-forward Perceptrons can't solve this problem is that doing so entails applying the XOR (exclusive or) computing function, which classifies whether a line segment is part of one contiguous shape in the image but not part of another. Yet a single layer of nodes without feedback is mathematically incapable of implementing XOR because it essentially has to classify all the data in one go with a linear rule (e.g., "If both of these nodes fire, the function output is true"), and XOR requires a feedback step ("If either of these nodes fires, *but they don't both fire*, the function output is true").

When Minsky and Papert reached this conclusion, it effectively killed most of the funding for the connectionism field, and it would be decades before it came back. But in fact, back in 1964 Rosenblatt explained to me that the Perceptron's inability to deal with invariance



was due to a lack of layers. If you took the output of a Perceptron and fed it back to another layer just like it, the output would be more general and, with repeated iterations of this process, would increasingly be able to deal with invariance. If you had enough layers and enough training data, it could deal with an amazing level of complexity. I asked him whether he had actually tried this, and he said no but that it was high on his research agenda. It was an amazing insight, but Rosenblatt died only seven years later, in 1971, before he got the chance to test his insights. It would be another decade before multiple layers were commonly used, and even then, many-layered networks required more computing power and training data than was practical. The tremendous surge in AI progress in recent years has resulted from the use of multiple neural net layers more than a half-century after Rosenblatt contemplated the idea.

So connectionist approaches to AI were largely ignored until the mid-2010s, when hardware advances finally unlocked their latent

potential. Finally it was cheap enough to marshal sufficient computational power and training examples for this method to excel. Between the publication of *Perceptrons* in 1969 and Minsky's death in 2016, computational price-performance (adjusting for inflation) increased by a factor of about 2.8 *billion*.[28] This changed the landscape for what approaches were possible in AI. When I spoke to Minsky near the end of his life, he expressed regret that *Perceptrons* had been so influential, as by then connectionism had recently become widely successful within the field.

Connectionism is thus a bit like the flying-machine inventions of Leonardo da Vinci—they were prescient ideas, but not workable until lighter and stronger materials could be developed.[29] Once the hardware caught up, vast connectionism, such as one-hundred-layer networks, became feasible. As a result, such systems were able to solve problems that had never been tackled before. This is the paradigm driving all the most spectacular advances of the past several years.

THE CEREBELLUM: A MODULAR STRUCTURE

To understand neural networks in the context of human intelligence, I propose a small detour: let's go back to the beginning of the universe. The initial movement of matter toward greater organization progressed *very* slowly, with no brains to guide it. (See the section "The Incredible Unlikeliness of Being," in chapter 3, regarding the likelihood of the universe to have the ability to encode useful information at all.) The amount of time needed to create a new level of detail was hundreds of millions to billions of years.[30]

Indeed, it took billions of years before a molecule could begin to formulate coded instructions to create a living being. There is some disagreement over the currently available evidence, but most scientists place the beginning of life on earth somewhere between 3.5 billion and 4.0 billion years ago.[31] The universe is an estimated 13.8 billion years old (or, more precisely, that's the amount of time that has passed

since the big bang), and the earth likely formed about 4.5 billion years ago.[32] So around 10 billion years passed between the first atoms forming and the first molecules (on earth) becoming capable of self-replication. Part of this lag may be explained by random chance—we don't know quite how unlikely it was for molecules randomly bumping around in early earth's "primordial soup" to combine in just the right way. Perhaps life could have started somewhat earlier, or maybe it would have been more likely for it to start much later. But before any of those necessary conditions were possible, whole stellar lifecycles had to play out as stars fused hydrogen into the heavier elements needed to sustain complex life.

According to scientists' best estimates, about 2.9 billion years then passed between the first life on earth and the first multicellular life.[33] Another 500 million years passed before animals walked on land, and 200 million more before the first mammals appeared.[34] Focusing on the brain, the length of time between the first development of primitive nerve nets and the emergence of the earliest centralized, tripartite brain was somewhere over 100 million years.[35] The first basic neocortex didn't appear for another 350 million to 400 million years, and it took another 200 million years or so for the modern human brain to evolve.[36]

All through this history, more sophisticated brains provided a marked evolutionary advantage. When animals competed for resources, the smarter ones often prevailed.[37] Intelligence evolved over a much shorter period than prior steps: millions of years, a distinct acceleration. The most notable change in the brains of pre-mammals was the region called the cerebellum. Human brains today actually have more neurons in the cerebellum than in the neocortex, which plays the biggest role in our higher-order functions.[38] The cerebellum is able to store and activate a large number of scripts that control motor tasks, such as one for signing your signature. (These scripts are often informally known as "muscle memory." This is not, in fact, a phenomenon of muscles themselves but rather of the cerebellum. As an action is repeated again and again, the brain adapts to make it easier and more