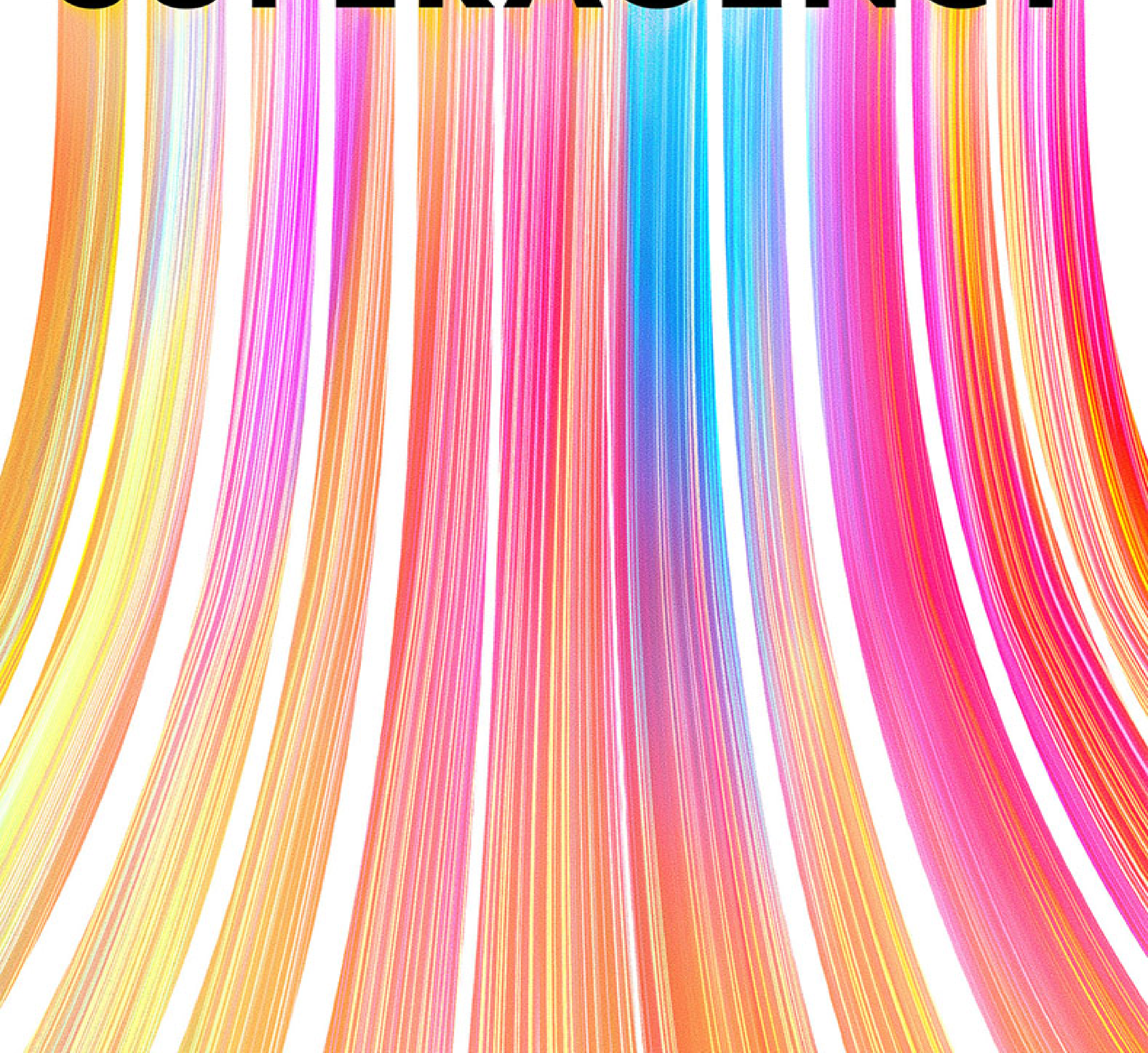
REID HOFFMAN
+ GREG BEATO

What Could Possibly Go Right with Our AI Future

SUPERAGENCY



SUPERAGENCY

What Could Possibly Go Right with Our AI Future

REID HOFFMAN
and
GREG BEATO

AUTHORS EQUITY

A NOTE ON THE TEXT

This book is a collaboration between my coauthor Greg Beato and me. We use “we” when representing our collective viewpoint. In instances specific to details from my own life, we revert to “I.”

—Reid

CONTENTS

INTRODUCTION

Throughout history, new technologies have regularly sparked visions of impending dehumanization and societal collapse. The printing press, the power loom, the telephone, the camera, and the automobile all faced significant skepticism and sometimes even violent opposition on their way to becoming mainstays of modern living.

Fifteenth-century doom-mongers argued that the printing press would dramatically destabilize society by enabling heresy and misinformation, and by undermining the authority of the clergy and scholars. The telephone was characterized as a device that could displace the intimacy of in-person visits and also make friends too transparent to one another.[1] In the early decades of the car's ascent, critics claimed it was destroying family life, with unmarried men choosing to save up for Model Ts instead of getting married and having kids, and married men resorting to divorce to escape the pressures of consumption that cars helped create.[2]

This same kind of doom and gloom was applied to society-wide automation in the 1950s, when increasingly sophisticated machines were dramatically impacting factories and office buildings alike, with everyone from bakers, meatcutters, autoworkers, and U.S. Census Bureau statisticians seeing their overall numbers dwindle. In 1961, *Time* magazine reported that labor experts believed that without intervention from business interests, unions, and the government, automation would continue to grow the "permanently unemployed."[3] By the mid-1960s, congressional subcommittees were regularly holding hearings regarding the mainframe computer's potential threat to privacy, free will, and the average citizen's capacity to make a life of their own choosing.

Today, U.S. unemployment rates are lower than they were in 1961. The average U.S. citizen lives in a world where PCs, the internet, and smartphones have ushered in a new age of individualism and self-determination rather than crushing authoritarian compliance or the end of humanity. But with the emergence and ongoing evolution of highly capable AIs, it's not just that familiar fears about technology persist; they're growing.

Even among AI developers, some believe that future instances of superintelligent AIs could represent an extinction-level threat to humanity. Others point out that, at the very least, humans acting with malicious intent will be able to use AIs to create catastrophic damage well before the machines themselves wage unilateral war against humanity. Additional concerns include massive job displacement, total human obsolescence, and a world where a tiny cabal of techno-elites capture whatever benefits, if any, AI enables.

The doomsday warnings are different this time, these observers insist, because the technology itself is different this time. AI can already *simulate* core aspects of human intelligence. Many researchers believe it will soon attain the capacity to act with complete and extremely capable autonomy, in ways that aren't aligned with human values or intentions.

Robots and other kinds of highly intelligent systems have long existed in sci-fi novels, comic books, and movies as our dark doppelgangers and adversaries. So as today's state-of-the-art AIs hold forth like benevolent but coolly rational grad students, it's only natural to see foreshadowing of HAL from *2001: A Space Odyssey*, or the Borg from *Star Trek*, or, in a less self-aware and more overtly menacing form, *The Terminator*'s relentless killer robot. These narratives have shaped our worst visions of the future for a long, long time.

But are they the right narratives? The future is notoriously hard to foresee accurately—for pessimists and optimists alike. We didn't get the permanent mass unemployment that labor experts in the early 1960s anticipated; nor did we get *The Jetsons* and its flying cars—at least not yet.

As hard as it may be to accurately predict the future, it's even harder to stop it. The world keeps changing. Simply trying to stop history by entrenching the status quo—through prohibitions, pauses, and other efforts to micro-manage who gets to do what—is not going to help us humans meet either the challenges or the opportunities that AI presents.

That’s because as much as collaboration defines us, competition does too. We form groups of all kinds, at all levels, to amplify our efforts, often deploying our collective power against other teams, other companies, other countries. Even within our own groups of like-minded allies, competition emerges, because of variations in values and goals. And each group and subgroup is generally adept at rationalizing self-interest in the name of the greater good.

Coordinating at a group level to ban, constrain, or even just contain a new technology is hard. Doing so at a state or national level is even harder. Coordinating globally is like herding cats—if cats were armed, tribal, and had different languages, different gods, and dreams for the future that went beyond their next meal.

Meanwhile, the more powerful the technology, the harder the coordination problem, and that means you’ll never get the future you want simply by prohibiting the future you *don’t* want. Refusing to actively shape the future never works, and that’s especially true now that the other side of the world is only just a few clicks away. Other actors have other futures in mind.

What should we do? Fundamentally, the surest way to prevent a bad future is to steer toward a better one that, by its existence, makes significantly worse outcomes harder to achieve.

At this point we know from thousands of years of experience that if a technology can be created, humans will create it. As I’ve written elsewhere, including in my previous book, *Impromptu*, we’re *Homo techne* at least as much as we’re *Homo sapiens*. We continuously create new tools to amplify our capabilities and shape the world to our liking. In turn, these tools end up shaping us as well. What this suggests is that humanism and technology, so often presented as oppositional forces, are in fact integrative ones. Every new technology we’ve invented—from language, to books, to the mobile phone—has defined, redefined, deepened, and expanded what it means to be human.

We’re the initiators of this process, but we can’t fully control it. Once set in motion, new technologies exert a gravity of their own: a world where steam power exists works differently than the world that preceded it. This is precisely why prohibition or constraint alone is never enough: they offer stasis and resistance at the very moment we should be pushing forward in pursuit of the brightest possible future.

Some might describe this as technological determinism, but we think of it as navigating with a kind of techno-humanist compass. A compass helps us to choose a course of action, but unlike a blueprint or some immutable manifesto, it's dynamic rather than determinative. It helps us orient, reorient, and *find* our way.

It's also crucial that this compass be explicitly humanist, because ultimately every major technological innovation impacts human agency—our ability to make choices and exert influence on our lives. A techno-humanist compass actively aims to point us toward paths in which the technologies we create broadly augment and amplify individual and collective agency.

With AI, this orientation is especially important. Because what happens to human agency when these systems and devices, often described as agents themselves, do become capable of replacing us entirely? Shouldn't we slow down that eventuality as much as possible? A techno-humanist perspective sees it the other way around: our sense of urgency needs to match the current speed of change. We can only succeed in prioritizing human agency by actively participating in how these technologies are defined and developed.

First and foremost, that means pursuing a future where billions of people around the world get equitable, hands-on access to experiment with these technologies themselves, in ways of their own choosing. It also means pursuing a future where the growing capabilities of AI help us reduce the threats of nuclear war, climate change, pandemics, resource depletion, and more.

In addition, it means pursuing this future even though we know we won't be able to predict or control every development or consequence that awaits us. No one can presume to know the exact final destination of the journey we're on or the specific contours of the terrain that exists there. The future isn't something that experts and regulators can meticulously design—it's something that society explores and discovers collectively. That's why it makes the most sense to learn as we go and to use our techno-humanist compass to course-correct along the way. In a nutshell, that's "iterative deployment," the term that OpenAI, ChatGPT's developer, uses to describe its own method in bringing its products into the world. It's a concept my coauthor, Greg Beato, and I will explore and emphasize throughout this book.

As a longtime founder and investor in technology companies, my perspective is inevitably shaped by the technology-driven progress and positive outcomes I've participated in over the course of my career. I was a founding board member at PayPal and part of its executive team when eBay purchased it in 2002. I cofounded LinkedIn and have sat on Microsoft's board since 2017, following its purchase of LinkedIn.

I was also one of the first philanthropic supporters of OpenAI when it launched as a nonprofit research lab in 2015. I led the first round of investment in 2019 when OpenAI established a for-profit limited partnership in order to support its ongoing development efforts. I served on its board from 2019 to early 2023. Along with Mustafa Suleyman, who cofounded DeepMind, I cofounded a public benefit corporation called Inflection AI in 2022 that has developed its own conversational agent, Pi. In my role at the venture capital firm Greylock, I've invested in other AI companies. On my podcast *Possible*, I regularly talk with a wide range of innovators about the impacts AI will have on their fields—with a techno-humanist compass guiding our conversations. I also provide philanthropic support to Stanford University's Institute for Human-Centered Artificial Intelligence (HAI) and to the Alan Turing Institute, the United Kingdom's national institute for data science and artificial intelligence.

I recognize that some might say such qualifications actually disqualify my perspective on AI. That my optimism is merely hype. That my idealism about how we might use AI to create broad new benefits for society is just an effort to generate economic return for myself. That my roles as founder, investor, advisor, and philanthropic supporter of many AI-focused companies and institutions create an ongoing incentive for me to overpromote the upsides and downplay the dangers and downsides.

I argue that the opposite is true: I'm deeply involved in this technology and I want to see it succeed exactly because I believe it can have profoundly positive impacts on humanity. My engagement in this domain has meant that I've seen firsthand the progress being made. That has strengthened my commitment, and thus I've continued to invest in and support a widening range of companies and organizations. I stay alert to potential dangers and downsides, and am ready to adapt, if necessary, precisely because I want this technology to succeed in ways that broadly benefit society.

One reason iterative deployment makes so much sense in the case of pioneering technologies like AI is that it favors flexibility over some grand master plan. It makes it easier to change pace, direction, and even strategy when new evidence signals the need for that.

Meanwhile, here we are presenting our argument to you in a book.

Roughly 2,400 years ago, Socrates critiqued the written word for its lack of dynamism in Plato's *Phaedrus* and for the way it made knowledge accessible to anyone:

> You know, Phaedrus, writing shares a strange feature with painting. The offsprings of painting stand there as if they are alive, but if anyone asks them anything, they remain most solemnly silent. The same is true of written words. You'd think they were speaking as if they had some understanding, but if you question anything that has been said because you want to learn more, it continues to signify just that very same thing forever. When it has once been written down, every discourse rolls about everywhere, reaching indiscriminately those with understanding no less than those who have no business with it, and it doesn't know to whom it should speak and to whom it should not.[4]

For Socrates, apparently, fixing his thoughts into written text represented a loss of agency. Had he turned his teachings into books himself, or rather scrolls, the reigning technology of his day, he would not have been able to control who read them. He would not have always been on hand to provide updates on his thinking, elaborate on nuances in the text, or correct misreadings. Consequently, face-to-face dialogic inquiry was his preferred technology for transmitting ideas.

But clearly generations of authors and readers thought differently. Why? Because ultimately written works increased the agency of authors and readers, enabling the latter to engage with, learn from, modify, expand upon, and, yes, perhaps even misinterpret or appropriate ideas from authors with which they might never have otherwise crossed paths.

As printing technologies improved, books evolved into a transformative global resource. Rolling about everywhere, indiscriminately reaching everyone, they functioned as early mobility machines, decoupling human cognition from human brains, democratizing knowledge, accelerating human progress, and providing a way for individuals and whole societies to benefit from the most profound and impactful human insights and innovations across time and space.

Of course, there are myriad other ways to share information now, and we'll be using many of them to convey the ideas in *Superagency* too. Along with the usual podcasts and social media, we'll be experimenting with AI-generated video, audio, and music to augment and amplify the key themes we're exploring here. To see how, check our website Superagency.ai.

But we're starting with a book—in part as homage to the essential truth that technologies that often seem decidedly flawed and even dehumanizing at first usually end up being exactly the opposite.

CHAPTER 1

HUMANITY HAS ENTERED THE CHAT

As 2022 drew to a close, people around the world continued to navigate the complex landscape of postpandemic recovery. Polls showed that inflation had replaced Covid-19 as the top global concern.[1] Food prices remained at record highs. But the return to normalcy was also in full swing. Job growth and average hourly wage increases for the month of November would greatly exceed estimates.[2] Tickets to Taylor Swift's concert tour were in such high demand that Ticketmaster's systems buckled under the weight of 14 million simultaneous customers.

Still, the tech industry itself was having some trouble shaking off the cumulative effects of advertising slowdowns, shifting investor sentiments, and evolving user engagement patterns. On November 9, Meta, the company formerly known as Facebook, laid off 11,000 employees, its largest-ever workforce reduction. Two days later, FTX, the Bahamas-based cryptocurrency exchange, declared bankruptcy with allegations of massive fraud and misuse of customer funds quickly following. The news rocked the entire cryptocurrency ecosystem, wiping out billions in market value. On November 14, the bad news continued, with the report that, like Meta, Amazon had begun to conduct unprecedented layoffs, with at least 10,000 employees to be let go in the coming weeks.

In part, these adverse outcomes were simply a correction to the surges in tech industry hiring, revenue, and market caps that pandemic stimulus and pent-up consumer demand had inspired. But they also provided a clear rebuttal to the ongoing narrative regarding Big Tech's alleged power to exercise complete control over markets and manipulate consumer behavior at will.

Throughout the 2010s, this narrative had effectively become gospel. Through addictive design practices and an arsenal of attention-hijacking techniques, it asserts, companies like Alphabet (aka Google) and Meta have all but perfected the dark art of maximizing engagement and trapping millions of users in digital hells of doomscrolling, rage-tweeting, shit-posting, rabbit-holing, thirst-trapping, hate-reading, group-shaming, and self-retweeting.

But none of Meta's allegedly irresistible tactics had managed to lure many people to the metaverse, its multibillion-dollar effort to create a vast virtual world where users could work, play, and socialize in immersive 3-D environments. And while Silicon Valley venture capitalists had been pouring billions into blockchain startups and other cryptocurrency projects, the developers of these technologies had not yet cracked the code for making decentralized finance as indispensable to mainstream users as the web, search, email, mobile computing, text messaging, and social media had quickly become in earlier waves of digital innovation.

Meanwhile, on the final day of November, after a month that had played out as one long swipe-left for the tech industry, OpenAI, a San Francisco–based research lab with a staff of around 375 employees, unveiled its latest product with no advance notice and zero hype. "Try talking with ChatGPT, our new AI system which is optimized for dialogue," the lab's official X.com account tweeted[3] at 10:02 a.m. "Your feedback will help us improve it."

OpenAI's cofounder and CEO, Sam Altman, was similarly circumspect in his first tweet[4] about ChatGPT: "language interfaces are going to be a big deal, i think. talk to the computer (voice or text) and get what you want, for increasingly complex definitions of 'want'! this is an early demo of what's possible (still a lot of limitations—it's very much a research release)."

As its name suggested, ChatGPT was a chatbot—not exactly many people's first pick to be the next *Pokémon Go*, or even the next Juicero. For most of their history on the web, where they were typically put to use in customer service contexts, chatbots functioned so poorly that the average person looking for guidance on how to return an online purchase still preferred to wait on hold for twenty minutes to speak to someone who sounded like they were answering from a call center located in the bottom of a swimming pool on Mars.

But ChatGPT was different. Very different. And that difference was immediately apparent. Impressively knowledgeable, stunningly versatile, and convincingly human, ChatGPT could provide an easy-to-understand explanation of quantum mechanics. It could compose sonnets about the Consumer Price Index, if that's what you wanted. It could help you debug—or possibly bug—your Python code. ChatGPT didn't always get things right, but even its mistakes, commonly described as "hallucinations," induced wonder and intrigue.

With zero marketing dollars behind it, ChatGPT attracted its first one million users in five days. Somehow, it seemed like two million of those one million people were journalists, reporting on their experiences—and that's when interest in ChatGPT really started to skyrocket. In just two months, it attracted 100 million users[5] and generated so much excitement, aspiration, and FOMO in the tech industry that it should have gotten a commendation from the Federal Trade Commission's Bureau of Competition.

Alphabet CEO Sundar Pichai sent a code-red alert to every Googler announcing that AI was now Job 1 for the entire company. Microsoft, which had invested in OpenAI three years earlier, started mentioning copilots more often than an airline training manual. Mark Zuckerberg announced that Meta had created a new top-level generative AI product group to "turbocharge" the company's efforts in the field.[6] Newcomers and upstarts like Anthropic, Midjourney, Hugging Face, and Replika pushed AI forward in ways that left the giants trying to keep up. Even research papers with titles along the lines of "Quantum Entanglement of Neural Networks in Multidimensional Latent Spaces" could go low-key viral on X.com.

In the face of these new conditions, sentiments took a 180-degree turn. For years prior to ChatGPT's release, Big Tech's biggest critics had been insisting that antitrust actions were necessary to inject new competitive energy into this once-dynamic U.S. technology sector. Now those same critics started saying that innovation was happening too quickly, that it was out of control. In March 2023, a nonprofit organization called the Future of Life Institute published an open letter that urged "all AI labs to immediately pause for at least 6 months the training of AI systems more powerful than GPT-4."[7] More than 33,000 people, including many AI-industry leaders and technologists, signed it. Their mood was dire, their urgency palpable. The Senate Judiciary Committee took note and conducted multiple hearings on AI oversight throughout the year.

Six months came and went, then another six months. Developers kept working, and the innovations continued. OpenAI continued to release updates to GPT-4, the foundation model underlying ChatGPT that was achieving state-of-the-art performance on complex tasks and problem-solving, and gaining the capacity to analyze images and provide feedback on them. Anthropic's Claude 2 achieved new levels of factual accuracy and expanded its context length, meaning it could process and keep track of context in input texts of up to about 75,000 words. If you wanted to summarize the complete and unabridged version of H. G. Wells's *The War of the Worlds* in twenty bullet points, Claude could do that.

Still, many of the ongoing challenges remained. Conversational agents like ChatGPT and Claude are built on top of large language models, or LLMs, a specific kind of machine learning construct designed for language-processing tasks. LLMs like GPT-4 process and generate language using what's known as neural network architecture, in which multiple layers of nodes perform a complex cascade of interconnected computations. Each node in a layer takes input from the previous layer, applies mathematical operations, and passes the result to the next layer. Parameters, as they're called, are also pivotal in this process, as they determine the strength of connections between nodes.

In a process known as pretraining, LLMs learn associations and correlations between *tokens*—words or fragments of words—by scanning a vast amount of text. In an LLM, each parameter functions something like a tuning knob, and in today's largest models, there are hundreds of billions of them. Through an iterative process of adjusting these parameters across all nodes in the network, the model reinforces or reduces connections between the tokens in its training data and begins to recognize and replicate complex patterns in language.

What this means, in part, is that LLMs never know a fact or understand a concept in the way that we do. Instead, every time you prompt an LLM with a question, or ask it to take some action, you are simply asking it to make a prediction about what tokens are most likely to follow the tokens that comprise your prompt in a contextually relevant way. And they don't always make correct or appropriate predictions.

By expanding pretraining datasets, fine-tuning model performance on more task-specific datasets, along with other measures, developers try to make their models more accurate and less prone toward undesirable outputs. As models grow more capable, they begin to display a sophisticated kind of simulated "awareness" of the world, such as recognizing that when a person says "I'm so hungry I could eat a horse" they're using hyperbolic language for expressive impact rather than asking for horse recipes.

But even when it seems like models possess humanlike commonsense reasoning, they don't. Instead they're making statistically probable predictions regarding patterns of language. This means they sometimes make mistakes. They can behave unpredictably. When models generate false information or misleading outcomes that do not accurately reflect the facts, patterns, or associations grounded in their training data, they are said to be "hallucinating"—that is, they're "seeing" something that isn't actually there.[*]

That means a model might provide an incorrect answer to a question that has a correct answer. It might fabricate entirely novel "facts," such as names, dates, or events, that have no basis in reality. It can provide information that may be accurate but has no contextual relevance to a given user prompt. Finally, it can generate outputs that are logically inconsistent or incoherent.

In addition, a model's dependence on data and quantification may give it the appearance of objectivity or neutrality, but it's not objective or neutral. Instead it's created by human developers and institutions making choices about which data to collect, how to process that data, what purpose or specific function a model is being optimized for, how best to align that function with human values and intents, and so on.

If a model's training data contains sexist or racist sentiments—which can happen when massive quantities of text are scraped from the internet and subjected to little or no additional filtering, vetting, or refinement—then the model might produce sexist or racist outputs. If a developer creating an AI for medical diagnosis doesn't fully grasp the complexities of certain medical conditions, that could lead to models that perform poorly for underrepresented patient groups or rare diseases.

Another issue involves the often opaque ways large language models operate, a characteristic known as the "black box" phenomenon. This occurs when complex neural networks processing hundreds of billions of text samples in extremely granular fashion identify patterns that human overseers have trouble discerning—

making it hard or even impossible to explain a model’s outputs or trace its decision-making process.

While developers apply various techniques to mitigate these issues, the fundamental limitation that underlies them all remains the same. As of yet, LLMs have no real capacity for commonsense reasoning, no lived experience, and no grounded model of the world. They’re always just predicting the next token in a sequence, based on patterns they’ve learned from their training data.

So even when built on top of a state-of-the-art language model, ChatGPT and its peers continue to hallucinate. They can still tie themselves in knots trying to solve relatively simple brain-teasers. They sometimes generate biased outputs in some instances and contextually irrelevant ones in others.

And this won’t ever completely change, skeptics assert. Even as developers build bigger supercomputers on which to train their models with trillions of parameters, and expand training datasets to include multimodal inputs like images, videos, and structured data, performance gains have started to slow. Errors persist. Critics say they’ll never achieve AI’s holy grail—*artificial general intelligence*, or AGI, in which models become capable of applying knowledge from one domain or context to entirely different situations, adapting to new challenges with humanlike flexibility, reasoning abstractly across diverse fields, and generating original ideas and solutions—all without being explicitly programmed for each task.

The Beginning Is Near

Throughout 2023, thanks to LLMs, AI was the biggest story in tech. Many observers believed these new models were on the verge of changing everything, in a good way. Many others believed they were on the verge of changing everything, in cataclysmically bad ways. Still others believed they were on the verge of keeping everything the same, only more so, in terms of concentrating power, profits, and the future of the world in the hands of a few Big Tech players.

By the summer of 2024, however, an ironic shift had occurred. Whereas generative AI critics had once demanded a six-month pause in the development of systems more powerful than GPT-4, for fear of potential catastrophic risks, there were now questions about why it was taking so long to deliver the next generation of groundbreaking models. And along with those questions there were persistent and increasing doubts regarding how much more capable LLMs might ultimately get. Thus, what had once been portrayed as Public Enemy No. 1 was now being deemed a dud. Phrases like “AI hype,” “AI bubble,” and “troughs of disillusionment” began to seep into media headlines.

But I’d already been through similarly dizzying changes in expectations—in the opposite direction. In 2015, when I first became involved with OpenAI, the idea that AI could eventually achieve, or credibly simulate, humanlike understanding and reasoning remained on the fringes of conventional wisdom. Even in Silicon Valley, the prospect was considered an extreme long shot. That was one of the reasons OpenAI had established itself as a nonprofit. Venture capital firms seeking a return on investment within the traditional VC time frame of five to ten years would have been unlikely to commit to such a speculative and long-term endeavor.

And if there were big challenges to overcome in 2024, well, there had been big challenges in 2015 too. There were big challenges in 2018 and 2020. Somehow OpenAI and the rest of the AI development community had always managed to find new techniques and breakthroughs that enabled the next wave of improvement.

Granted, it’s easy to be an optimist if your time horizons are long. And mine are. In fact, I believe that we’re still in the very early stages of this new phase of human discovery and growth. The supercomputers are going to get even more super. Developers will continue to write more efficient algorithms. To overcome some of the limitations that characterize LLMs, they’ll come up with new architectures and techniques, and incorporate different approaches like multimodal learning and neurosymbolic AI—systems that integrate neural networks with symbolic reasoning based on explicit, human-defined rules and logic.

What all that means is that we’re also still in the early stages of the existential reckoning that these systems will provoke, as we try to process what it really means to introduce new, and not entirely predictable, forms of intelligence into our world.

A machine that can think like a human—strategically, abstractly, and even creatively, at the speed and scale of a computer—will obviously be revolutionary. What if every child on the planet suddenly has access to an AI tutor that is as smart as Leonardo da Vinci and as empathetic as Big Bird? What if billions of people around the world suddenly have a highly knowledgeable and reliable health care advisor in their pocket at all times?

Of course, not everyone focuses on the potential upsides of AI—especially in the U.S., where people consistently express high levels of concern about the technology.

According to a 2022 survey conducted by Ipsos, a global analytics firm, “only 35% of sampled Americans (among the lowest of surveyed countries) agreed that products and services using AI had more benefits than drawbacks.”[8] A similar survey from the Pew Research Center found that only 15 percent of U.S. adults said they were “more excited than concerned about the increasing use of AI in daily life.”[9] In another study, from Monmouth University, 56 percent said that “artificially intelligent machines would hurt humans’ overall quality of life.”[10]

Such concerns are entirely understandable. We’re experiencing a world-changing moment that’s creating significant uncertainty. Exactly how good will these systems get, and how fast? What kinds of jobs will be left for people as AI continues to improve? What happens to trust and public discourse, already under siege, as AI technologies make it cheaper and easier to produce convincing simulations of reality at scale? What happens to individual privacy and autonomy in a highly instrumented world, where billions of systems, devices, and robots can match or exceed human performance, and can take actions that may infringe on our own choices and desires? Can we continue to maintain control of our lives, and successfully plot our own destinies?

Most Concerns About AI Are Concerns About Human Agency

It’s this last question that informs this book’s primary focus: Can we continue to maintain control of our lives, and successfully plot our own destinies? Ultimately, questions about job displacement are questions about individual human agency: Will I have the economic means to support myself, and opportunities to engage in pursuits I find meaningful? Questions about disinformation and misinformation are questions about individual human agency: How do I know whom and what to trust as I make decisions that impact my life? Questions about privacy are questions about individual human agency: How do I maintain the integrity of my own identity and how I’m known in the world, and preserve an authentic sense of self?

Human agency is a fundamental concept in philosophy, sociology, and psychology. It holds that you, as an individual, have the capacity to make your own choices, act independently, and thus exert influence over your life. While you may also believe that external circumstances and conditions play a significant role in the