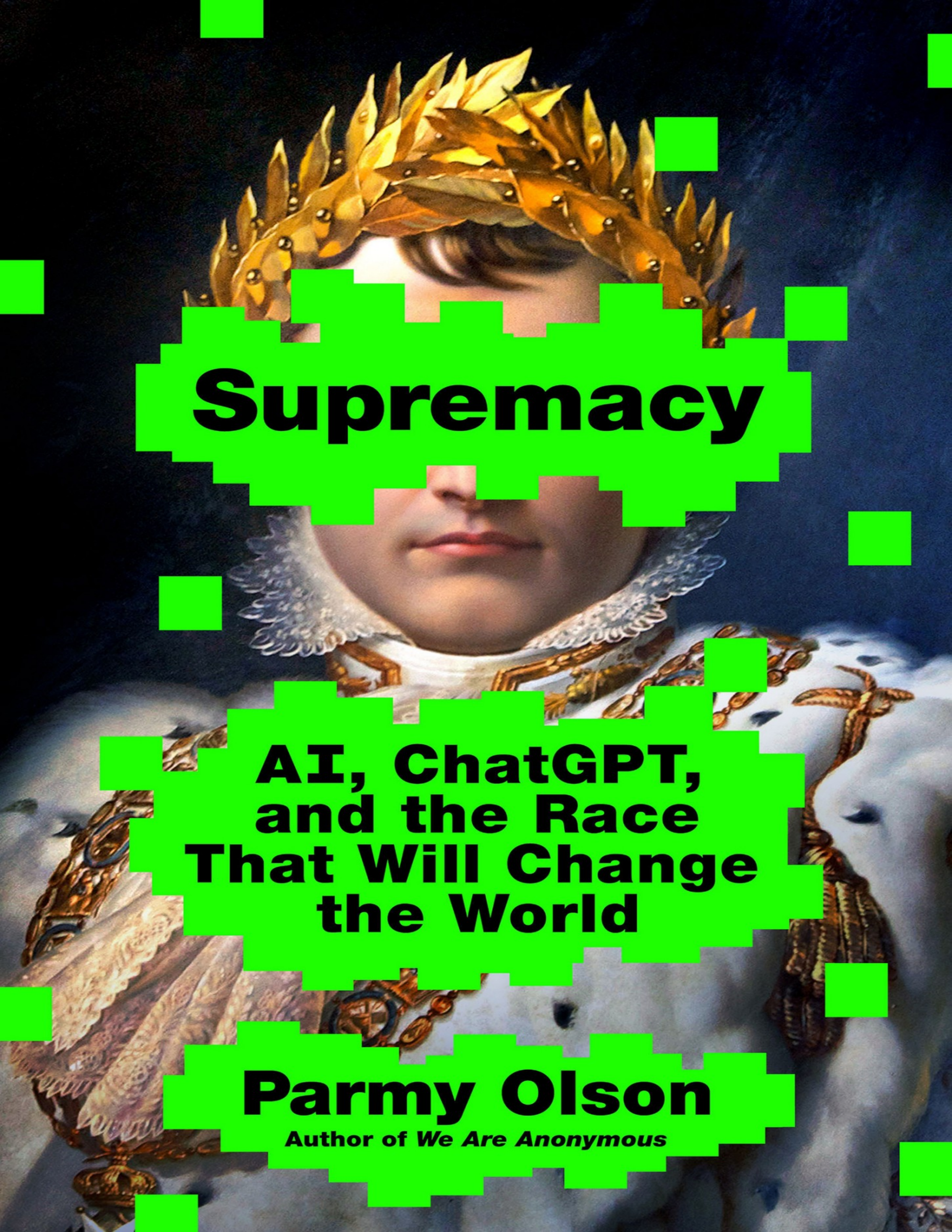

Supremacy

AI, ChatGPT, and the Race That Will Change the World

Parmy Olson

Author of *We Are Anonymous*

PARMY OLSON

SUPREMACY

AI, ChatGPT, and the Race That Will Change the World

St. Martin's Press New York

Begin Reading

Table of Contents

About the Author

Copyright Page

To Mani

PROLOGUE

After picking up this book and reading these first few words, you might be wondering if a human wrote them.

That’s OK. I’m not offended.

Two years ago, that thought would not have even crossed your mind. But today, machines are generating articles, books, illustrations, and computer code that seem indistinguishable from the content created by people. Remember the “novel-writing machine” in the dystopian future of George Orwell’s *1984* and his “versificator” that wrote popular music? Those things exist now, and the change happened so fast that it’s given the public whiplash, leaving us wondering whether today’s office workers will have jobs in the next year or two. Millions of white-collar professionals suddenly look vulnerable. Talented young illustrators are wondering if they should bother going to art school.

What’s remarkable is how quickly this has all come to pass. In the fifteen years that I’ve written about the technology industry, I’ve never seen a field move as quickly as artificial intelligence has in just the last two years. The release of ChatGPT in November 2022 sparked a race to create a whole new kind of AI that didn’t just process information but *generated* it. Back then, AI tools could produce wonky images of dogs. Now they are churning out photorealistic pictures of Donald Trump, whose pores and skin texture look so lifelike they’re almost impossible to distinguish as fake.

Many AI builders say this technology promises a path to utopia. Others say it could bring about the collapse of our civilization. In reality, the science fiction scenarios have distracted us from the more insidious ways AI is threatening to harm society by perpetuating racism, threatening entire creative industries, and more.

Behind this invisible force are companies that have grabbed control of AI’s development and raced to make it more powerful. Driven by an insatiable hunger to grow, they’ve cut corners and misled the public about their products, putting themselves on course to become highly questionable stewards of AI.

No other organizations in history have amassed so much power or touched so many people as today’s tech giants. Google conducts web searches for 90 percent of Earth’s internet users, and Microsoft software is used by 70 percent of humans with a computer. But neither company is satisfied. Microsoft wants to grab a chunk of Google’s $150 billion search business, and Google wants Microsoft’s $110 billion cloud business. To fight their war, each company has grabbed the ideas of others—which is why, when you boil everything down, our AI future has been written by just two men: Sam Altman and Demis Hassabis.

One is a scrawny but placid entrepreneur in his late thirties who wears sneakers to the office. The other is a former chess champion in his late forties who is obsessed with games. Both are fiercely intelligent, charming leaders who sketched out visions of omnipotent AI so inspiring that people followed them with cult-like devotion. Both got here because they were obsessed with winning. Altman was the reason the world got ChatGPT. Hassabis was the reason we got it so quickly. Their journey has not only defined today's race but also the challenges coming our way, including a daunting struggle to steer AI's ethical future when it is under the control of industry giants.

Hassabis risked scientific ridicule when he established DeepMind, the first company in the world intent on building AI that was as smart as a human being. He wanted to make scientific discoveries about the origins of life, the nature of reality, and cures for disease. "Solve intelligence, and then solve everything else," he said.

A few years later, Altman started OpenAI to try to build the same thing but with a greater focus on bringing economic abundance to humanity, increasing material wealth, and helping "us all live better lives," he tells me. "This can be the greatest tool humans have yet created, and let each of us do things far outside the realm of the possible."

Their plans were more ambitious than even the craziest Silicon Valley visionaries. They planned to build AI that was so powerful it could transform society and make the fields of economics and finance obsolete. And Altman and Hassabis alone would be the purveyors of its gifts.

In their quest to build what could become humankind's last invention, both men grappled with how such transformative technology should be controlled. At first they believed that tech monoliths like Google and Microsoft shouldn't steer it outright, because they prioritized profit over humanity's well-being. So for years and on opposite sides of the Atlantic Ocean, they both fumbled in the dark for novel ways to structure their research labs to protect AI and make benevolence its priority. They promised to be AI's careful custodians.

But both also wanted to be first. To build the most powerful software in history, they needed money and computing power, and their best source was Silicon Valley. Over time, both Altman and Hassabis decided they needed the tech giants after all. As their efforts to create superintelligent AI became more successful and as strange new ideologies buffeted them from different directions, they compromised their noble goals. They handed over control to companies who rushed to sell AI tools to the public with virtually no oversight from regulators, and with far-reaching consequences. The concentration of power in AI would lead to reduced competition and herald new intrusions into private life and new forms of racial and gender prejudice. Already today, if you ask a popular AI tool to generate images of women, it'll make them sexy and scantily clad; ask it for photorealistic CEOs, and it'll generate images of white men; ask for a criminal, and it will often generate images of Black men. Such tools are being woven into our media feeds, smartphones, and justice systems, without due care for how they might shape public opinion.

The pair's journey was not all that different from one two centuries ago, when two entrepreneurs named Thomas Edison and George Westinghouse went to war. Each had pursued a dream of creating a dominant system for delivering electricity to millions of consumers. Both were inventors-turned-entrepreneurs, and both understood that their technology would one day power the modern world. The question was this: Whose version of the technology would come out on top? In the end, Westinghouse's more efficient electrical standard became the most popular in the world. But he didn't win the so-called War of the Currents. General Electric did.

As corporate interests pushed Altman and Hassabis to unleash bigger and more powerful models, it was the tech titans who came out as the winners, only this time the race was to replicate our own intelligence. Now the world has been thrown into a tailspin. Generative AI promises to make people more productive and bring more useful information to our fingertips through tools like ChatGPT. But every innovation has a price to pay. Businesses and governments are adjusting to a new reality where the distinction between real and "AI-generated" is a crapshoot. Companies are throwing money at AI software to help displace their employees and boost profit margins. And a new breed of personal AI devices that can conduct an unimaginable new level of personal surveillance is cropping up.

The second half of this book lays out those risks, but first I'll explain how we got here, and how the visions of two innovators who tried to build AI for good were eventually ground down by the forces of monopoly. Their story is one of idealism but also one of naivety and ego, and of how it can be virtually impossible to keep an ethical code in the bubbles of Big Tech and Silicon Valley. Altman and Hassabis tied themselves into knots over the stewardship of AI, knowing that the world needed to manage the technology responsibly if we were to stop it from causing irreversible harm. But they couldn't forge AI with godlike power without the resources of the world's largest tech firms. With the goal of enhancing human life, they would end up empowering those companies, leaving humanity's welfare and future caught in a battle for corporate supremacy. This was how it happened.

ACT 1

THE DREAM

CHAPTER 1

High School Hero

Sam Altman knew he ought to keep his mouth shut. In the conservative stronghold of St. Louis, Missouri, people didn’t talk about whether they were gay or straight. While the rest of America was coming to grips with gay rights, Altman’s midwestern hometown was lagging in the early 2000s and still made it a crime to sleep with someone of the same gender. Teenagers like him who had an inkling they were gay tended to find safety in silence. Altman was different. He had to speak up, not because he wanted people to know everything about him, but because talking about it would become a mission.

Altman was that one kid in high school who seemed to magically transcend the labels that others tried to slap on him. He was as bright as any geek, charismatic as any jock. In his English literature assignments, he’d emulate the challenging prose of Faulkner, and in math, he breezed through calculus. Then he’d jump into the pool to bark orders to his water polo team, which he captained, or head home to coordinate video games with his friends for hours on end. At the dinner table with his younger brothers, Max and Jack, he’d geek out about space travel and rocket ships, and then when they played a board game, like Samurai, Sam would declare himself the leader. In this and many other situations, he liked to take charge.

Altman grew up in a middle-class Jewish family, his mother, Connie, a dermatologist and his father, Jerry, a lawyer. Jerry helped push for affordable housing in the city of St. Louis, as well as the reconstruction of historic buildings, and his actions fueled his son’s public-spirited view of the world. Sam vividly remembers Jerry bringing him to his office one day and telling him that even if he didn’t have the time to help somebody, “you have to figure out how to do it anyway,” he says.

Sam also had an enormous amount of self-belief that came from being the eldest of four children as well as a brazenness that others respected. He talked openly about his sexuality when other kids his age, and kids generally in the late nineties, would have kept it secret. He leaned into something that many folks in the Midwest believed was bad and made it seem cool, in part because he wanted to help others who were like him.

That calling came from the internet. When Altman started logging onto the web portal America Online (AOL), he realized there were many more out there like him. Here was the wonderful thing about logging onto AOL’s chat rooms. You’d hear the dial tone and beeps denoting the “handshake” as your modem negotiated a secure connection to the World Wide Web, and the screeching tonal overtures of what sounded like a busted CB radio. Then you got connected, your heart beating a little faster at all the possibilities before you, all the chat rooms. Here you could talk to another interesting person on a computer in another part of the world. The rooms had names like "Beach Party" or "The Breakfast Club." Some of the biggest rooms were noisy and teeming with creeps, but if you explored rooms in more specific categories like Pet Lovers, X-Files Fans, or Gay & Lesbian, you'd find more coherent conversations.

For people like Altman, these rooms became a lifeline. You could hide under the anonymity of a screen name and lurk while others talked about LGBTQ-friendly places to go. They gave him a sense of belonging in a world where it was easy to feel like an outsider. "Finding AOL chat rooms was transformative," he would say later in a *New Yorker* magazine profile. "Secrets are bad when you're eleven or twelve."

AOL chat rooms were so significant to the LGBTQ community that by 1999, when Altman was fourteen, about a third of its rooms were focused on gay topics. When he was sixteen, he came out to his parents. His mother was shocked. Her son had always seemed "unisexual and tech-y," she later said in that same magazine profile on her son, but he also tended not to fit classifications. For instance, he was a vegetarian in a part of America where everybody loved barbecue. He was obsessed with computers, but he also wasn't reclusive or socially awkward. And while everyone else was listening to nineties pop, he preferred classical music.

The Altmans transferred their precocious teen to John Burroughs, an elite private school on the outskirts of St. Louis that boasted a sprawling, leafy campus and sought to grow its students' talents for "the improvement of human society."

He stepped up to as many leadership roles as he could manage. As well as captaining the water polo team, he edited the yearbook and spoke at school assemblies. He socialized with the faculty while occasionally stretching the rules so he could make a stir. At the annual fall pep rally, Altman and his water polo team ripped off their clothes on stage so they were left standing in their Speedos, grinning to whoops and cheers.

He got in trouble for that with the school's athletic director, but rather than let it go and complain about the admonishment to his friends, or even to another teacher, he fought back and went straight to the top. He knocked on the door of school principal Andy Abbott, a former English teacher with a gentle demeanor. The older educator was charmed by the gangly, dark-haired teen with eyes like saucers, who frequently dropped into his office to pitch ideas or complain about an injustice that he planned to write about in the student paper.

The young man was utterly unintimidated by authority, he noticed. If Principal Abbott ever made an unpopular decision that affected the other students, the kid would take it upon himself to be their savior. "He would object," Principal Abbot remembers, "and thoughtfully." To this day, the soft-spoken educator sees Altman as "the most brilliant kid I knew."

That same earnestness, along with an ability to come across as candid and vulnerable, would later help the young man curry favor with other powerful figures in the worlds of technology and government, from investors to the press, to some of the world's most powerful CEOs, his solemn stare imploring them to support an epic mission. Altman would learn over time that people in power could smooth the road for your ambitions, just as Principal Abbott would do for him in high school.

His big school project was to build a real-life version of the network he'd found on AOL. He pushed through the school's red tape and got his principal's blessing to create the school's first LGBTQ support group. It was like an underground network that students could visit for counseling or to meet others like themselves. Within a year, about a dozen students had joined.

But Altman wasn't satisfied. He started approaching his teachers, asking them to put stickers on their doors that said their classrooms were a safe space for gay students, trying to turn his teachers into allies. He eventually started a Gay-Straight Alliance group, aimed at raising public awareness of gay rights.

He then decided to make a splash at the morning assembly. His new group went into the main hall early and placed a series of printed numbers on all the chairs before the audience took their seats. As Altman went up to the mic, he asked everyone with a certain one of the numbers to stand up. "Look around you," he told the audience, as about sixty kids stood up. "That's one in ten of you. That's how many people in the school identify as gay."

It was a bold demonstration, but something was odd. Several students were missing from the audience, and they all happened to be part of the school's Christian club. Altman found out later that they had boycotted his presentation by staying at home or in their classrooms. Indignant that these kids had pushed back against his objective, he once again marched into the office of Principal Abbott, demanding that the Christian kids be counted as absent.

"There was nothing damaging about making them more aware," the teenager argued. He wasn't a table thumper, but it was clear by his words and his stern expression that he was angry.

"I tried to justify it for a while," Principal Abbott remembers. "But I think he was probably right about that."

Altman left school with a hard lesson. If you had ambitious ideas, there would always be some haters. The solution was to align yourself with those who had power and authority and to surround yourself with a support network.

Soon enough, Altman got accepted into the prestigious Stanford University in the heart of California's Silicon Valley, a fount of brilliant software engineers and technology entrepreneurs who populated the sun-drenched region with technology start-ups. Despite his interest in programming and his admittance to a computer science degree, the lanky eighteen-year-old couldn't stand to focus on one subject alone. He was fascinated by everything. He took an array of humanities classes and creative writing classes.

Then outside of school hours, he'd drive twenty minutes south for lessons that would become critical to his future life as a globally renowned entrepreneur. He'd play hours of poker at a popular casino in San Jose, honing his skills of psychological maneuvering and influence. Poker is all about watching others and sometimes misdirecting them about the strength of your hand, and Altman became so good at bluffing and reading his opponents' subtle cues that he used his winnings to fund most of his living expenses as a college student. "I would have done it for free," he would later tell one podcast. "I loved it so much. I strongly recommend it as a way to learn about the world and business and psychology."

The field that Altman would one day focus on to transform the world came up as part of his regular degree. He became a researcher at Stanford's AI lab, a pocket of the university's vast campus that was filled with cables and the odd robot arm. The AI lab had just been reopened and its leader was Sebastian Thrun, a computer scientist with radical views, a soft German lilt, and piercing blue eyes. Thrun was part of a new breed of academics who weren't content to spend their days writing grant proposals and waiting for tenure but who worked with tech giants. Stanford was just five miles from Google's headquarters, and Thrun also ran the cutting-edge "moonshot" projects at Google X that made self-driving cars and augmented reality glasses.

In class, Thrun taught his students about machine learning, a technique that computers used to infer concepts from being shown lots of data instead of being programmed to do something specific. The concept was critical in the field of AI, even though the term *learning* was misleading: machines can't think and learn as humans do. Thrun noticed that the serious kid from St. Louis was interested in the possibility of unintended consequences in AI. What would happen if a machine learned to do the wrong thing?

Thrun explained that AI systems could act in unpredictable ways to achieve their "fitness function," or goal. If an AI was designed with a fitness function to survive and reproduce, it might inadvertently wipe out all biological life on Earth, Thrun said. This didn't mean the AI was bad. It was just unaware of the gravity of what it was doing. Its motives weren't all that different from ours when we washed our hands. We didn't hate the bacteria on our skin and want to destroy it. We just wanted clean hands.

Altman mused on this idea for some time. As a science fiction fan, he wondered if this was why humans had never had contact with alien life. Perhaps beings on other planets had tried creating AI, too, and then been wiped out by their own creation. If that was avoidable, someone would have built safer AI before others created the dangerous kind.

This seed of an idea would lay dormant in the back of Altman's mind for a little over a decade before blossoming into the creation of OpenAI. But for now it was far too big to tackle. Academics like Thrun built AI systems. Stanford students like Altman built start-ups that became companies like Google, Cisco, and Yahoo. The young geek wanted to do the same. He just needed a business idea. Then it came to him when he was walking out of class. “Wouldn’t it be great if I could open my mobile phone and see a map of where all my friends are?” he asked his Stanford classmate and friend Nick Sivo.

What if he created a digital map for his mobile phone on which he could find his friends and make that the main product of a company? Starting a company wasn’t easy. You needed to raise money from venture capitalists, and while there were dozens of those within three miles of Stanford, Altman was young and inexperienced. His answer came from the other side of the United States, in Cambridge, Massachusetts, where an older tech mogul was starting what he called a boot camp for young entrepreneurs. Altman and Sivo decided to join the three-month program, called Y Combinator, and create a start-up. Y Combinator would go on to become the most successful start-up accelerator of all time, seeding $400 billion worth of tech companies with names like Airbnb, Stripe, and Dropbox.

Altman, who was now nineteen, didn’t know any of this at the time. Most investors in Silicon Valley dismissed Y Combinator as a silly summer camp for hackers. Its creator was Paul Graham, a forty-one-year-old cargo shorts–wearing computer scientist who’d become a millionaire after selling his e-commerce company to Yahoo. Having achieved his wealth, Graham had fashioned himself as a thought leader, publishing essays on his website about topics beyond the purview of software geeks, from economics to having kids, from free speech to being a nerd in high school.

But his most popular essays, the ones that made kids like Altman sit up in the kind of rapt attention given to spiritual leaders, were about building start-ups. Over and over, those essays emphasized that the quality of a start-up founder mattered more than anything else. You didn’t need a brilliant idea to start a successful tech company. You just needed a brilliant person behind the wheel.

“Google’s plan, for example, was simply to create a search site that didn’t suck,” Graham wrote. And look where that had led. Lightbulb moments were passé. It was the founders who mattered, and the best were hackers—programmers who were willing to break conventional wisdom to build new things. As a hacker, “you could be 36 times more productive than you’re expected to be in a random corporate job,” he wrote.

Starting a tech company in Silicon Valley was even patriotic because it emulated the rugged individualism of America’s founding fathers: “Hackers are unruly,” he wrote. “That is the essence of hacking. And it is also the essence of Americanness. It is no accident that Silicon Valley is in America, and not France, Germany, England, or Japan. In those countries, people color inside the lines.”

The path was almost simple, Graham taught. Bootstrap your company, start with a minimum viable product, and optimize it over time. Work in a tight bubble, because it was better to have ten people love what you made than thousands liking it. And don’t be afraid to bend the rules along the way. In fact, why not rewrite society itself?

Graham's ideas would eventually strike a huge chord in Silicon Valley and fuel a prevailing new wisdom that the vision of a start-up's creators was so sacrosanct that they should be allowed to act with impunity, like gods. It's why the founders of Google and Facebook could fashion themselves as modern-day autocrats of business, often grabbing a majority of a company's voting shares and sometimes taking their companies in strange directions. (A case in point was the lack of pushback from Facebook's board or shareholders to Mark Zuckerberg's strange and expensive decision to become a virtual reality company.) Thanks to something called a dual-class share structure, many tech start-up founders, including those behind Airbnb and Snapchat, could hold these unusual levels of control of their companies. Graham and others believed founders had this authority for good reason. When the smartest and most talented people had a long-term vision, they needed the freedom to carry it out.

Graham saw those same hacker instincts in Altman: deeply curious, fiercely intelligent, and a big thinker. And there was something else. This teenager with unruly dark hair was comfortable around older people to the point that he'd have no trouble managing people like Graham, who was twenty years his senior. When Graham suggested Altman wait a year to join his Y Combinator program since he was only nineteen, the kid replied that he was coming anyway. Graham liked him immediately.

Most of the other entrants to the program were engineers and hackers, including the founders of the popular online forum Reddit. Graham and his wife, Jessica Livingston, gave $6,000 to each start-up in their new program, a number based on the stipend that MIT gave grad students during the summer. While most venture capitalists threw millions of dollars at start-ups, Graham told founders to do more with less, and aim for "ramen profitability," discouraging them from hiring lawyers, bankers, and PR people so they could do that work more cheaply themselves.

Graham himself did everything on a shoestring. He cooked dinner every Tuesday night, his specialty being chicken fricassee, and brought in his friends as speakers to talk about start-ups, while Livingston took care of the legal paperwork for each new company.

Altman and his friend Sivo named their new company Loopt, and he moved to Cambridge, Massachusetts, to work out of Y Combinator's first office near Graham's house. Altman and Graham formed a close relationship similar to the one he'd had with his high school principal. Among the hopeful young entrepreneurs, Graham was like a spiritual leader whom they just called PG. Altman took Graham's teachings seriously and saw Loopt not as a company that could make him rich but as an idea that could make the world a better place. He tinkered nonstop on the prototype, living off instant ramen noodles and Starbucks coffee ice cream. He worked so hard and ate so badly that he developed scurvy, a disease caused by a lack of vitamin C.

Though he was a decent enough programmer, the boyish-faced Altman was an even better businessman. He had no qualms about calling up executives from Sprint, Verizon, and Boost Mobile and pitching a grand vision about changing the way people socialized and used their phones. Speaking in low tones and using elegant turns of phrase that he'd honed from his creative writing classes, he explained that Loopt would one day be essential to anyone who had a mobile. App stores didn't exist yet, so he had to rely on mobile operators to preinstall Loopt on some of the earliest smartphones. That's why getting the telecom executives on board was so key, and Altman was a master at marketing his new company. Sprint, Verizon, Boost, and even BlackBerry agreed to put his service on their phones.

At the end of Y Combinator's three-month program, Altman raised some money so he could grow his start-up. He presented the Loopt vision to a group of fifteen investors, most of whom were Graham's wealthy friends, for about fifteen minutes and then reached out to people with even deeper pockets: venture capital firms in Silicon Valley. He got several offers and took $5 million from two premier venture capital firms who'd backed Google, Yahoo, and PayPal.

With all that funding, Altman dropped out of Stanford University to work on Loopt full-time. He moved with a handful of engineers he'd hired to Palo Alto, California, settling into Sequoia's coworking space. They were churning out code late into the night alongside the creators of YouTube before Altman moved his team to their first office in Mountain View, prime real estate that was a few blocks away from the headquarters of Google. It was as close to the heart of Silicon Valley as you could get.

Silicon Valley was the land of crazy thinkers. You didn't start a business here. You started an empire. Or you sought to build something on the frontier of technology and science. If you wanted to conduct scientific research on a disease like Alzheimer's, you could go to a university on the East Coast or in Europe. But if you wanted to reverse aging, you went to Silicon Valley.

The region's network was its big selling point. On any given day you could bump into someone at an event who could juice your business. Go to breakfast at Buck's in Woodside, California, and you might catch the cofounder of Yahoo having a fruit and yogurt combo while sitting at the same table where Elon Musk had his first funding meeting for PayPal. Grab a drink at the Musto Bar at the Battery Club in San Francisco and you could spot one of the cofounders of Facebook.

Altman quickly integrated himself into Silicon Valley's web of programmers, investors, and executives. If you knew how to plug yourself into this modern old-boys network, you were more likely to get swept up in the success that propelled its members to billionaires. Altman was so good at networking that he managed to make the right connections to present Loopt at Apple's prestigious annual conference for developers in 2008. Dressed in jeans and two green and pink polo shirts that gave him the look of a children's TV presenter, the slender young entrepreneur told his audience that Loopt was the largest social mapping service in the world. "We make serendipity happen," he said, staring at the crowd and barely cracking a smile.

On the surface, everything looked perfect. But underneath, Loopt was struggling. People just weren't all that excited about using Altman's digital map to find their friends. The wide-eyed entrepreneur had believed that young mobile users wanted to meet their friends just as much as he did. But the premise of helping you meet your buddies at a bar or connecting to total strangers when you needed an extra player at basketball was a lot of effort when you could just interact with them through a screen. As the aughts wore on, more people were engaging with their friends on social networks like Facebook. Facebook was growing considerably faster than Loopt was. It had racked up hundreds of millions of active users, while Loopt had barely managed to get five million people just to register.

It also didn't help that Loopt was becoming controversial. A year after he founded the company, Altman got a call from his old high school principal, Andy Abbott. Parents were forcing their kids to use Loopt so they could track their whereabouts, Abbott said, and at one point a parent had called the school during a field trip to say their child's bus was speeding. "Look what you've wrought," Altman's old mentor said on the phone, half joking.

Altman had heard worse. "We're getting concerns from women's groups," the young entrepreneur admitted. Some men were making their wives install Loopt so they could follow their whereabouts at all times. It was a creepy and potentially dangerous way to misuse Altman's creation. "But we're working on a solution," he quickly added. Loopt's users could fake their location. A vulnerable woman who was at the grocery store could make it look like she was at home instead.

While many entrepreneurs would have been in denial about their app's misuse, Altman seemed intent on openly confronting the problem. He'd learned as a teenager that keeping things secret just made them worse. It was better to bring them out in the open. He got a call from Jessica Lessin, then a dogged tech reporter with the *Wall Street Journal*, asking about Loopt's privacy issues and some of the concerns of misuse. To Lessin's surprise, Altman was eager to talk about the controversy, so much so that he emailed her a long document detailing all the risks that came with using his app, according to an account she later published.

What looked like career suicide was a shrewd PR move that he would turn to again and again in the future as a kind of calculated reverse psychology. By becoming overly concerned with the worst-case scenario of his creation, Altman could disarm his critics or journalists like Lessin. There was nothing left to throw at him because, well, he'd done it to himself. He seemed almost too honorable for his own good—even though the honorable thing to do with an app that people used to stalk the vulnerable might have been to shut it down.

In the end, consumers did that for him. Altman had miscalculated how uncomfortable they felt about pinging their GPS coordinates to meet up with others. "I learned you can't make humans do something they don't want to do," he would go on to say.

The wiry young entrepreneur had spent most of his twenties frantically trying to grow Loopt to no end. He'd used the iPhone's new push notifications to ping people's home screens to get them using Loopt's chat feature. He helped advertisers send "flash deals" to Loopt's users. He talked up each upgrade like it was a slam-dunk success. "The response has been tremendous," he said in one 2010 interview. But it was mostly hot air. By 2012, only a few thousand people around the world were using Loopt regularly. So much for building an empire. Like the vast majority of technology start-ups, Loopt had failed.

In the world of tech start-ups, a founder's ultimate goal is to either have their idea become a multibillion-dollar company or make a multibillion-dollar exit by selling their firm to a bigger fish. It was becoming harder to stay independent, and most were getting swallowed up by a tech giant like Google or Facebook. If a founder could do that, they'd often use the spoils to start a new firm and kick-start the cycle all over again as a serial entrepreneur. But Loopt's exit wasn't exactly spectacular. In 2012, Altman sold it to a gift-card company for about $43 million, barely covering what was owed to investors and his employees.

He could have ditched Silicon Valley then and there, but Loopt's collapse emboldened him with a greater conviction that he should do something more meaningful. He wasn't the first tech maverick to find bigger ambitions in the ashes of failure. About a decade earlier, Elon Musk had been ousted from PayPal by his board. Smarting from the experience, Musk decided he was tired of working on something as superficial as a consumer payments service. "The next company I do [should] have some long-term beneficial effect," he told an interviewer. A few years later he did just that. Musk met the founders of Tesla and worked on saving humanity from the existential threat of climate change.

If you throw a smartphone at a group of people sitting inside the exclusive Battery Club in San Francisco, it'll hit at least three trying to save the world. Many a Silicon Valley entrepreneur has believed their app was elevating humanity, and while some have indeed created useful products used by millions of people, many others have also developed a full-blown messiah complex as a result. The region's emphasis on innovation made this savior culture pervasive, fueled by Graham's founder-knows-best principles. If you were among that top breed of innovative hackers, you could solve not only engineering problems but also societal conundrums that had beset humanity for years.

Altman wanted Loopt to bring people together because that's what they needed. We were all becoming glued to our screens, mindlessly scrolling and sprinkling "likes" across different social networks to create an increasingly quantified sense of human connection. He needed to try something more meaningful. Maybe Altman had to give people what they didn't know they wanted. That was what Apple had been doing successfully for years, and in Silicon Valley, that was the secret everybody was trying to crack.

The young St. Louis native needed to delve back into the world of building start-ups and plug himself so deeply into those Silicon Valley networks that he would become synonymous with the companies who proclaimed they were changing the world. He would transform himself into an even more profound version of his old mentors, and then dig back into what he’d been stewing on at the Stanford AI lab. That would lead him to chase an even grander objective: saving humanity from a looming existential threat and then bringing them an abundance of wealth unlike anything they had seen.

CHAPTER 2

Winning, Winning, Winning

The sounds of screams, the roar of a roller coaster, and a clanking fairground organ marked the start of the 1994 computer game *Theme Park*. A large square of pixelated grass, empty, was waiting to be filled with food stalls in the shape of giant hamburgers and roller-coaster tracks that shot up to nosebleed heights. The goal was to make as much profit as possible.

Theme Park wasn't made by middle-aged game designers eager to teach business principles to kids, but by a dark-haired teenager from North London named Demis Hassabis (pronounced hah-SAH-bis). He had the work ethic of a Silicon Valley entrepreneur and was obsessed with playing games. Years before Hassabis would become the front-runner in a race to build the world's smartest AI systems, he was learning how to run a business via simulation, something that would become a running theme in his life's work and in his quest to build machines more intelligent than humans.

In *Theme Park* you started with a cash balance of about $200,000 to pay for the construction of rides and staff wages. You earned that back from the sale of things like tickets, merchandise, ice cream, and coconut shy games. If you didn't hire enough mechanics, the rides would break down; not enough security guards, thugs would overrun the park. Skimp on sugar and visitors would balk at the ice cream. Staff would go on strike, and wages got negotiated. Though he was just seventeen, Hassabis designed this tricky balancing act of cost and reward as a highly complex imitation of business management, and he made the game so addictive that it sold fifteen million copies after its release in 1994.

Video games had hit the UK and the United States like a tidal wave, pulling kids into vivid dopamine worlds where Teenage Mutant Ninja Turtles fought their way through side-scrolling levels or where you could steer a pickup truck around wild dirt tracks. But Hassabis thought the best video games were simulations that acted as microcosms of real life. God games let you wield the power of creation and destruction. Instead of controlling a single character like Mario, you shaped the lives of thousands of virtual characters, fashioning landscapes or directing the progress of a civilization. You could build a city and then hit it with a natural disaster, or you could fill a theme park with hundreds of visitors.

You could have fun with this technology, but you could also learn things, from how to run a business to the mysteries of the universe. For all the entertainment value of games, Hassabis would eventually become gripped by a powerful desire to use them to create an artificial superintelligence that would help him unlock the secrets of human consciousness.

That calling to understand the mysteries of the universe went beyond the aims of most other scientists, which could seem incongruous until you think about how Hassabis grew up an enigma himself, the lone mathematical genius in a family of bohemian creatives. His mother, Angela, was a devout Baptist who immigrated to the UK from Singapore, then met her future husband in the host family she was staying with in North London, a free-spirited Greek Cypriot named Costas Hassabis. The two seemed as mismatched as socks and sandals, but they married and had three children, starting with Demis. Costas bounced around between different jobs, like teaching and running a toy shop, and moved his family ten times before Hassabis was twelve years old.

By then, it was clear that Demis was different from other kids. He'd been beating his father and uncle at chess at the age of four, and by six, he was thrashing most kids his age at local chess championships, tottering on a cushion or telephone book to help him peer over the board. He was a fluent reader and curious about everything, but Hassabis channeled most of his brainpower into games. When his dad brought home board games with missing pieces, he'd use them to design new ones and play them with his younger brother and sister.

But the real fun was yet to come. A decade before Sam Altman found AOL chat rooms in the 1990s, Hassabis was diving headfirst into an even more rudimentary kind of technology, made up of chunky pixels on plain black screens. In 1984, when he was eight, Hassabis used some money he'd won at a chess competition to buy a ZX Spectrum 48. One of the earliest personal computers, it consisted of a thick black keyboard that you hooked up to your TV that used cassette tapes to play colorful graphics on the screen.

Hassabis bought programming books and taught himself to create games for the Spectrum. Before going to bed, he'd set up a calculation to run overnight and then go to sleep. The next morning, the calculations were done. For Hassabis, this was a revelation. He'd offloaded his cognitive labor to the Spectrum. The computer was acting like an extension of his mind.

He got hooked on the niche world of programming and upgraded to a more powerful Commodore Amiga 500, a chunky white collection of gadgets that included a mouse and its monitor, and started a hacking club with his school friends. They'd write code to conjure colorful snippets on the screen that copied a scene from a game they'd played, and Hassabis was often driven to outdo his friends by making things more intricate. He'd take his computer apart. Then put it back together. He created a digital chess game and got his younger brother, George, to play it.

Chess was still the focus of his life, and Hassabis wanted to become a world champion. His mother supported his aspirations and started homeschooling him so he could spend more time studying the game. During school vacations, he was traveling to different chess tournaments to compete, and the work was relentless. Hassabis would later say that games were like a gym for the mind and chess was the ultimate workout. Just as poker taught Sam Altman about psychology and business, chess taught Hassabis how to strategize by starting with the end in mind. You visualized a goal and worked backward.

But everything changed one day when Hassabis was eleven and competing in a chess championship in Liechtenstein. He was playing against Denmark's national chess champion, and the match had become a marathon. After ten hours of play, their brains fighting exhaustion, the Dane tried to force a draw. The young Hassabis still had his king and queen, but his opponent had the advantage with his king, rook, bishop, and knight. Hassabis was tired and assumed he was about to be checkmated. He resigned.

The Danish champion was shocked. "Why did you resign?" he asked.

He showed Hassabis that he could have done a drawing move instead. Hassabis stared at the board. Sometimes failure sparks bigger ambitions. If you couldn't stand to lose, chasing a bigger prize could be a consolation, and Hassabis had just bombed after an enormous effort. As he looked around the room at the other chess geniuses hunched over their boards, neurons churning, it dawned on him that this whole tournament was a waste of brainpower. These were some of the world's top strategic thinkers. What if they worked on solving bigger problems? He was now the second-best chess player in the world under the age of fourteen, but this was still just a game.

Hassabis told his parents he wanted to stop going to chess tournaments and enrolled back into school. He was a quiet, sentimental kid, who listened to Enya and taught himself to play her song "Watermark" on the piano. His favorite movie was *Blade Runner*, a science fiction flick about a detective who hunts down rogue AI replicants that are nearly indistinguishable from human beings, and he found himself wrapped up in the film's most emotive moments. He'd play the stirring Vangelis soundtrack from its final scene over and over, when the villain laments that his memories will be "lost in time, like tears in rain."

On Sundays, his mother regularly took Hassabis and his siblings to North London's Hendon Baptist Church, a grand, gray-stoned structure that looked over the suburbs from atop a hill. It was a small pocket of internationalism, its worshippers coming from places like the Philippines, Ghana, France, and India. For a half-Cypriot, half-Singaporean kid like Hassabis, this was a place to blend in. Sundays were a livelier alternative to the buttoned-up services of the dominant Church of England. People would raise their hands in praise and sing to the beat of the pounding drum kit and band. The pastor glossed over dogma to emphasize treating others with respect. Prayers were emotional, and the church itself was unashamedly evangelical.

Baptists were minnows in the UK, despite being one of the biggest Christian denominations in America. They had about 150,000 members compared to the one million affiliated with the Church of England. But religion and the concept of God fascinated Hassabis, and he wondered if God could be found through scientific means.

When he was sixteen, having just finished high school two years early, he read a book by Nobel Prize–winning physicist Steven Weinberg called *Dreams of a Final Theory*. It was about a grand, almost quixotic quest for a unifying theory of nature. Weinberg believed there might be a way to explain all the fundamental forces of the universe in a single set of equations, similar to how Einstein's equation $E = mc^2$ summed up the relationship between energy and mass. Ideally, this "theory of everything" would be succinct enough to be written on a single page or even as a single equation.

Hassabis was enthralled, and then shocked, to learn that scientists hadn't been making much progress on finding this framework. They needed help, he thought. They needed intellectual horsepower. Maybe he could help? Hassabis looked over at the chunky Commodore Amiga that had worked out calculations through the night while he slept. Maybe a smarter computer could help. If he could make computers more intelligent, a more capable extension of his mind, perhaps they could help scientists crack those hard questions about the universe, even discover a divine origin.

"It seemed like the perfect, mega solution," Hassabis later explained in an interview with *New York Times* writer Ezra Klein. He had thought of studying physics at university, but after reading Weinberg's book, believed he should go for something bigger. If he studied computer science and the burgeoning field of artificial intelligence, he could build the ultimate scientific tool and make discoveries that improved the human condition. Hassabis couldn't shake his fixation with games, and so he formed a long-term plan to marry both efforts. The key was to focus on games that simulated the real world. Games in the late 1980s were already able to simulate the rudimentary features of an entire civilization. If a computer could replicate the full Technicolor details of a world, perhaps a highly intelligent computer could figure out how to fix the real world's glitches, too—figure out what to do in the simulation and apply it to real life.

Hassabis drew inspiration from god games, and his favorite was *Populous*. "What fascinated me about them was that they were living worlds, and the game evolved with how you played it," he says. "You could simulate a part of the world as a sandbox and you could play around it."

The game's retro pixelated graphics belied its complexity. As a player, you were the deity of a plain green valley dotted with houses and had divine powers to lead their inhabitants—your followers—into battle against those of other deities. You could raise and lower the land elevation, making it flatter so your followers could build houses and breed more. You could cause earthquakes. *Populous* pioneered the god-game genre, and Hassabis loved the game so much that he became determined to work at the company that designed it, Bullfrog Productions. He entered a competition to win a job at the company and lost. So he picked up the phone and called the firm instead, asking for a week's work experience. They agreed and liked him enough to bring him back for a summer job when he was fifteen.

Soon after, when Hassabis secured a place at Cambridge to begin a computer science degree at the age of sixteen, the university told him he was too young and should wait at least a year. So he spent that time at Bullfrog again, getting paid in cash and staying at the local YMCA hostel near its office in Guildford, Surrey. Hassabis started as a video game tester and quickly got the job of level designer directly under Bullfrog's founder, Peter Molyneux.

With his cue-ball head and black polo shirts, Molyneux looked more like a pub landlord than a gaming legend. Over the years, he'd also become a divisive figure in the industry thanks to a widely reported habit of overpromising on projects. He'd make ambitious claims about game mechanics or features, saying, for instance, that players could plant an acorn in the virtual world of his game *Fable*, and then come back a few days later to find it had grown into a tree. That last part wasn't true.

But Molyneux also had big ideas that were driving the industry forward, and for the moment, he was basking in the glow of *Populous*. To the older entrepreneur, Hassabis came across as unusually curious, even precocious. The teen prodigy peppered his new boss with questions about the technical limits of Bullfrog's games and asked why they were calling certain features "artificial intelligence" when it seemed like they were basic software systems, Molyneux remembers.

"No task, however ridiculously big, seemed like a barrier to him," he says. When the Bullfrog founder wanted to build a simulation game about a theme park and all his other staff declined interest, preferring games with swords and fighting, Hassabis volunteered to build what would become the vivid world of roller coasters and food stalls in *Theme Park*. Molyneux became a mentor to Hassabis as the two designed the new game together, with a couple of digital artists supporting them. In between writing code and designing the gameplay, they often talked about the possibilities of artificial intelligence. Hassabis told his boss that he believed AI was probably a decade away from surpassing humans and reaching sentience.

"The future of AI felt like you could almost touch it," the senior game designer remembers. "The other question we often philosophized about was, 'Why should humans be the only ones that create things?' Why can't we have the burden of creativity handled by a piece of AI?" They imagined AI eventually writing music and poetry and even designing games.

For now, though, they were using systems that barely fit the description of AI to give *Theme Park* a touch of realism. Machine-learning techniques allowed them to give background characters their own personalities, with some of the park visitors more impulsive and prone to spending money and others tighter with their finances. The game was a hit. While *Populous* sold five million copies, *Theme Park* went on to sell three times more.

That made Hassabis something of a celebrity when he finally arrived at Cambridge. He borrowed Molyneux's Porsche 911 and drove it around campus, eager to impress some of the students there. After filling all his previous school holidays with chess tournaments, he treated his first year at university almost like a vacation, going out at night with his classmates and then lying in bed the next morning listening to The Prodigy while the sun streamed in through the windows. When he wasn't drinking red wine at the college bar, sending his cheeks bright red, he was playing speed chess or drag racing the borrowed Porsche. Eventually he crashed it and had to call his mentor to apologize. "It was the second time he crashed it," Molyneux remembers, cringing. But it was hard to be angry at the often-smiling whiz kid. "He's so charming."

Hassabis met members of his future inner circle at Cambridge, including Ben Coppin, another computer science student who would go on to lead product development at DeepMind, and with whom he talked about religion and how AI could solve global problems. But DeepMind was still more than a decade away. First, he had to graduate, and then go straight back to working for Molyneux. It was then that he stumbled upon the weirdest computer gaming job application he had ever seen. In the mail one day arrived a bottle containing a letter that its sender had smudged with tea stains. They had burned the edges and written a long missive in calligraphy, saying that they were shipwrecked on an island called Korporate. Hassabis understood the sentiment immediately because he would have hated to find himself toiling inside a giant corporation too.

The sender was Joe McDonagh, a programmer who worked at a sprawling conglomerate called British Telecom and who loved games. Joe was desperate to work for a gaming company, and to his delight, he was told to come to Molyneux's house for an interview. When he arrived, a small, elfin-looking young man with chin stubble and black hair that sat on his head like a helmet answered the door. It was Hassabis, and he looked much younger than his twenty-one years. "I thought, 'Who's this kid?'" McDonagh remembers. It turned out Hassabis was now an executive at the gaming firm and his interviewer.

McDonagh soon realized he had come face-to-face with another game enthusiast who was ferociously competitive. When McDonagh mentioned that he liked origami, Hassabis challenged him to a race to see who could fold a paper crane the fastest. Hassabis won. The pair played board games for the rest of the afternoon. Later, when McDonagh called back about the job, he was told that the unusual kid from his interview had left the company. Hassabis had left because his mentor could no longer keep up with the young man's technical aspirations. "We weren't moving fast enough for him," Molyneux remembers.

McDonagh managed to track down a phone number and called Hassabis to find out what was going on. "I'm setting up a new company," he explained. It was going to be called Elixir Studios, and it would have real, cutting-edge AI at the center of a god game that would eventually be used to simulate the world.

It was an extraordinarily ambitious vision. McDonagh was in. He joined Elixir and became a lead designer, a job that involved dreaming up extraordinary new worlds. Having learned the art of bombastic marketing from his old mentor, Hassabis was brazen and cocky about his goals when he spoke to the press. He was featured on the cover of *Edge*, a taste-defining magazine for computer games in the nineties, and boasted that he was going to build games that not only had superior capabilities but would drag gaming itself out of its niche market for teens. He would build something so clever that readers of *The Economist* would want to play it. "I wanted to show games could be a serious medium, like books and film," Hassabis says. He sketched out a long-term plan. Once Elixir was successful, he'd sell it and start an AI company.

He focused on building a flagship game called *Republic: The Revolution*, a political simulation where players had to overthrow the government of a fictional, totalitarian country in eastern Europe. Hassabis wanted everything to be as true to life as possible. McDonagh spent hours in the British Library studying the history of the Soviet Union to support his eager young boss in creating a realistic story. Hassabis worked more on the technical side, overseeing the creation of an artificial intelligence technique that could put a million virtual people in their game, an ambitious target considering that until then, the limit for most god games was closer to one or two thousand virtual background characters. He wanted players to be able to zoom in from a satellite image of a city down to a petal on the balcony of a tower block.

The former chess champion hired the smartest programmers he could find, many of them graduates from Oxford University and Cambridge. He stoked team spirit by giving them every opportunity to play games. He excelled at all of them, from the video game *Starcraft* to the strategic board game Diplomacy. Foosball was like a blood sport. Hassabis would use his signature Viper Shot at the table, using the whole of his arm to spin his plastic players and whack the ball into the goal. On the real pitch, he played striker for Elixir's team of out-of-shape software engineers. When they played soccer against a group of local North London guys in regular five-a-side matches, Hassabis was ferocious, attacking the ball and shins of other players twice his height like an angry terrier and, frequently, scoring.

As the team got closer to their deadline for the game, the boyish-faced founder and his programmers were working from 10:00 a.m. to 6:00 a.m. each day, stealing three or four hours of sleep in their board room, sometimes snoring at their desks, gamepads dangling from their hands. There were no late-night outings, and Hassabis himself made it a point to not let himself get drunk anymore, lest he impair his brain in any way.

His ambitions for *Republic*'s graphics and AI technology were bordering on absurd. He was building something that was lightyears ahead of the current computing capabilities. But there was no point in creating a virtual country if he couldn't populate it with thousands of real, breathing people. "I didn't want people to be abstract dots, wandering randomly across the player's screen," Hassabis told *Edge* magazine. "I wanted husbands, students, house-wives and drunks, each living separate, plausible lives."

There was no better way to showcase the magical capabilities of AI than through a game. At the time, the most advanced AI research was happening in the gaming industry as smarter software helped create living worlds and a new style called emergent gameplay. Instead of playing a set route in something like *Super Mario Bros*, you were plonked in the middle of a virtual world, given some tools, and left to fend for yourself. That was the essence of *Grand Theft Auto* and then *Minecraft*, which would become the best-selling video game in history.

Hassabis believed he was on track to do something similar—but there was a problem. *Republic: The Revolution* was boring. It was the worst possible trap to fall into for a game designer. The team had spent four of their five years of development time so focused on the technology that they had neglected to perfect the gameplay itself. Making a great computer game requires constant iteration. You generally have to start with something crude but playable and then play it thousands of times until it gets better. But Elixir's game makers couldn't spend the time they needed making the game more exciting because their boss's expectations for the technology had been so high.

"The essence of a video game is immersion and feeling," McDonagh remembers. "*Republic* didn't have either of those things. We got stuck in a tech black hole." The programmers at Elixir knew the game wasn't good enough. When it came out, the critics confirmed their worst suspicions and gave it mixed reviews, saying it was too complex. Sales were modest.

"It was too ambitious for its time," Hassabis admits. "I was in a rush to make a technical and artistic statement."

That didn't stop him from getting Elixir to try its hand again and launch another similarly ambitious god game called *Evil Genius*, in which players took the role of a James Bond–style villain that tried to achieve world domination. The game had some clever, tongue-in-cheek humor, but it also struggled to find mainstream success. As he took a stab at improving things with *Evil Genius 2*, Hassabis was facing huge costs from all his investments in the technology. In 2005, the brainy kid who wanted to stretch the limits of gaming found himself closing down Elixir. The experience crushed him. From chess to foosball to school, he'd spent his life winning at practically everything.

Blowback from the British gaming industry made the humiliation worse. He had masterfully built up excitement about Elixir in the press and the gaming industry as an upstart that would transform the old ways of gaming with new technology. McDonagh remembers attending a conference at one point and mentioning that he'd worked at Elixir. A grandee of the British gaming industry who overheard simply laughed, and McDonagh died inside. "Failure was incredibly hard," he remembers.

At one point, McDonagh and Hassabis got into a shouting match over the company's demise, the first and only time he heard the normally calm entrepreneur raise his voice. "It was miserable," he says. "We were all from Oxford and Cambridge. You're winning, winning, winning. We'd never lost before, and we lost so publicly."

In his passion to show people the magic of AI, Hassabis had made a crucial error in trying to build a game around his real passion. If he was going to make machines that were smarter than humans, he had to flip that strategy around. He needed to go deeper into artificial intelligence, and instead of using it to make a great game, use games to make great AI.

Several years later and now in his thirties, McDonagh found himself once again on the phone with his old boss, fielding a new job offer. “I’m starting this company called DeepMind,” said the eager entrepreneur who seemed to be a glutton for punishment.

I can’t do this again, McDonagh thought. He said, “No.”

He then watched in astonishment as Hassabis chased another impossibly ambitious goal, this time surpassing expectations to build what looked like the world’s most advanced artificial intelligence systems—until Sam Altman came along.

CHAPTER 3

Save the Humans

On a hot summer's day in 2006, Altman was lying on the floor of his studio apartment in Mountain View, California, in just his gym shorts. His arms were outstretched, and he was trying to breathe. He was halfway through a marathon weekend of trying to negotiate a deal for Loopt, and it wasn't going well. It was also ninety-five degrees inside. Altman felt like he was going to explode with stress, according to an account he shared with the Art of Accomplishment podcast in 2022.

He'd been telling himself for years that this was a normal part of being an entrepreneur. "This is how I'm supposed to feel," he thought. "But it isn't helping." The stress was making things worse.

The failure of Loopt had taught Altman that he couldn't force people to do what they didn't want to do. It also revealed a more personal lesson: how to emotionally disengage himself from difficult situations. That moment on the floor in his gym shorts became a turning point. Altman was going to live differently. The trick was to become more detached.

After selling Loopt and then breaking up with Nick Sivo, who had also been his long-term romantic partner, and then working for his acquiring company for a spell, Altman spent a year doing whatever he wanted. He detached himself completely. It generally doesn't look good to take a year off in the hustle-culture world of Silicon Valley, and Altman noticed the consequences immediately. If he started talking to someone at a party and mentioned his plans to take a year off, their eyes would look for someone else to talk to.

He kept a thread of a connection with California's Bay Area, working as a part-time partner with Y Combinator. By now, venture capital investors had changed their view of the scrappy hacker camp and saw it as a factory for high-quality internet companies. Several of its entrants had become big names like Reddit and Scribd. For start-up founders, "YC" was now seen as a gateway to success in the Valley. Thousands of tech founders applied and only about one hundred got through each year.

For the rest of Altman's self-imposed gap year, he dabbled in his wide array of interests, reading dozens of books about everything from nuclear engineering to synthetic biology and from investing to artificial intelligence. He traveled to other countries and stayed in hostels, flew to conferences and invested in several start-ups with some of the roughly $5 million he'd made from selling Loopt.

He would admit publicly that nearly all the companies he backed failed, but he figured he was training a muscle for identifying the projects that were most likely to succeed. It was OK to be frequently wrong, he believed, so long as you were occasionally "right in a big way," such as by backing a start-up that turned out to be a blockbuster and then making a spectacular exit.

If living life was like painting, Altman approached it with the biggest paint roller to cover the broadest area possible. But in dribs and drabs, he was increasingly drawn to artificial intelligence. At around the time he sold Loopt, he went on a hike with some tech industry friends and got into a discussion about the future of artificial intelligence research, according to his profile in the *New Yorker*. Altman considered that with computer hardware becoming so powerful, and machine learning systems so capable, at some point during his lifetime they would probably be able to replicate his brain.

This told him something important about humanity's role at the top of Earth's food chain. If our intelligence could be simulated by a computer, were we all that unique? Altman's answer to that question was no, and while that might have been a depressing realization at first, he flipped it to see it as something to capitalize on. If humans weren't so special, that meant they could be replicated by computers, even improved on. Maybe *he* could do that.

In many ways, Altman was building off a Silicon Valley mindset that saw life itself as an engineering conundrum. You could solve all manner of big problems by using the same steps you took to optimize an app. Partly this came from the way engineers were trained to approach technical problems systematically and logically, an approach that was deeply ingrained through their education and taught in software development. Success was measured in how efficient you could make your software. These prized methods naturally extended to other parts of society and life.

Little wonder Altman liked to use the language of computing when he referred to humans, such as when he once said in a magazine interview that "we learn only two bits a second." A bit is a basic unit of information, typically represented as a 0 or a 1 in binary code, and this was Altman's figurative way of showing how limited humans were in processing information. If you compared the mechanics of our brains to how computers worked, computers could process bits at a much more impressive rate: in gigabits or terabits per second.

If Altman himself was going to build a machine that surpassed human intelligence, there was no question that he would have to remain in Silicon Valley, where everybody was building for tomorrow.

"There is a relentless belief in the future here," he once said. "There are people here who will take your wild ideas seriously instead of mocking you." Silicon Valley also promised a thriving network of contacts who would scratch your back if you scratched theirs. Help someone fundraise for their start-up, and they might help you hire that talented engineer.

As his travels drew to a close, Altman started an early-stage investment firm called Hydrazine Capital to make financial bets on start-ups, from life sciences to education software companies. He drew on his connections with some of Silicon Valley's most powerful financiers. Paul Graham and Peter Thiel, an early investor in Facebook, added to the $21 million pot of money Altman raised for the fund. Thiel, known today as an enigmatic billionaire whose ideas border on science fiction, would become a kingmaker in the quest to build powerful AI, helping to fund both Altman and Demis Hassabis in London. He was part of the so-called PayPal Mafia, an elite group of cofounders and executives from the online payment giant who invested in one another's companies over many years and which included Elon Musk and LinkedIn founder Reid Hoffman.

Altman put about 75 percent of Hydrazine's money into companies graduating from Y Combinator, a strategy that paid off. Within about four years, the fund's value had grown tenfold, thanks to his investments in start-ups that were part of his growing network of contacts, many of them part of the Valley elite. He put money into Reddit, the start-up in his first YC class, and Asana, the enterprise software company started by Facebook's billionaire cofounder Dustin Moskovitz. In the coming years, both of those relationships would prove valuable in helping Altman build ultra-powerful AI.

Altman understood that immediate financial rewards weren't as valuable in the long run as personal connections. That's why he felt uncomfortable with the adversarial way he had to behave toward entrepreneurs as a venture capitalist. It was a job where you had to push for as much equity as possible, for as little money as possible. Altman also found Silicon Valley's constant striving for extreme wealth slightly distasteful. He was more interested in the glory that came from building exciting projects. In between his investing work, he pared down his assets to a four-bedroom house in San Francisco, a property in Big Sur, California, and $10 million in cash. He lived off its interest.

Then one day in 2014, Graham asked Altman a question while in the older entrepreneur's kitchen: "Do you want to take over YC?" Altman grinned. Graham and his wife, Jessica Livingston, had two young kids and were strung out from running a program that had become huge. For one thing, Graham tended to put his foot in his mouth when he gave interviews, often reinforcing suspicions that the Valley elite was dominated by white brogrammers. He once wrote in a blog post that he'd be "reluctant to start a startup with a woman who had small children, or was likely to have them soon."

As Graham's one-man show was starting to become a liability, Y Combinator was becoming more and more unwieldy. Over the previous seven years, it had funded 632 start-ups and was getting ten thousand applicants annually, allowing in just two hundred. More tech start-ups were being created than ever before, and Y Combinator needed to grow to meet the demand.

"I'm no good at running a giant thing," Graham said on stage at a conference later that year, explaining the leadership transition. "Sam is going to be good at running a giant thing."

Altman was just thirty at the time, and Graham was pushing fifty. But Altman was already acting like the new Graham. He had become his own version of a start-up guru with insights and advice on all manner of topics, including those with which he had little experience. Despite his youth and having only ever run one company that arguably failed, for instance, he once wrote a blog post with ninety-five nuggets of advice that other start-ups should follow.

However unseasoned Altman was, he'd made such a strong impression on Graham and Livingston that they never bothered to make a list of possible new leaders for YC. They both agreed it should just be Altman. Graham helped cast an almost messianic glow on his protégé, writing in an essay that "Sama," Altman's screen name, was one of the five most interesting founders of all time. "In questions of design, I ask 'What would Steve [Jobs] do?' but on questions of strategy or ambition I ask 'What would Sama do?'"

Once Altman took the reins of YC, his top priority was going bigger and, naturally, broader. He worked on turning the program into more of an institution, creating a board of overseers that included Jessica Livingston, Altman himself, and seven Y Combinator alumni. He doubled its full-time partners and added a handful more part-time partners, including Thiel, the billionaire venture capitalist.

Altman, who'd been interested in frontier science since he was a kid, believed that its advancements were critical to helping humanity and building wealth. So he focused on bringing in more "hard-tech" start-ups that were solving complex scientific and engineering problems. "It's just what I like doing," he remembers today. "And I don't really mind losing money in pursuit of something I think is worthwhile. I think it's important that we tackle our biggest challenges. While they require taking bigger risks, any potential returns are commensurate with that."

Till then, YC had mostly accepted consumer app makers and enterprise software companies who had more predictable paths to revenue. But Altman didn't think these were the companies that would transform the world. So he persuaded the founders of Cruise, a self-driving car start-up, to join the YC program, as well as Helion Energy, a nuclear fusion start-up based in Redmond, Washington.

Nuclear fusion happens when two light atomic nuclei combine to form a heavier nuclease, releasing energy in the process. It's the same reaction that powers the sun and the stars as well as the flux capacitor in *Back to the Future*'s DeLorean time machine and the arc reactor in Tony Stark's Iron Man suit. It's been a longtime holy grail for scientists looking for clean energy solutions, but it's also been decades away from reality. Most research in the field has led to theories and proofs of concept. But Helion, which was started by four academics, said it could build a nuclear fusion reactor for tens of millions of dollars instead of tens of billions and pave a path for people to transition away from burning fossil fuels.

It sounded crazy, but these were the kinds of big, world-changing ideas that Altman jumped at. He had been wanting to start his own nuclear energy company for years, and now he could invest in one instead.

He knew he was going against the grain of tech investing, which focused on software companies with more traditional business models and clearer paths to profit. But he firmly believed these companies could improve humanity and make a lot of money too. “Shame on Silicon Valley for not funding this company yet,” he said in one interview about Helion. Altman’s moral hubris wasn’t that odd. It was just a slightly different ideological flavor to that driving other Big Tech leaders, like Elon Musk, who was even more explicit about his goals of saving humankind.

“Another mobile app? You get an eye roll,” Altman once said. “A rocket company? Everyone wants to go to space.” In Silicon Valley, everyone also professes to want to save the world. But Altman was, like Musk, fashioning himself as a bona fide tech savior who was serious about his goals. Most tech entrepreneurs shared an implicit understanding that rescuing humanity was mostly a marketing ploy for the public and their employees, especially since their firms were building widgets that helped streamline email or do laundry. But Altman was refashioning the YC fraternity into a bigger, more serious alliance of entrepreneurs who really could fix the world. These were riskier bets that attracted more attention.

When it came to investing, Altman was the guy at a poker game who put most of his chips on the table for a half-decent hand and who raised the pulse rates of everybody else around the table. Altman attributed that tendency to a missing circuit in his brain—the one that made him care what people thought of him. It allowed him to calculate risk more effectively and to bet on investments that looked crazy.

When he did make those bets, though, Altman was cushioned from their failure by his wealth and his reputation as the new Yoda among start-ups. In Silicon Valley, a good reputation was more valuable than any mansion or sports car. And if you backed nuclear fusion start-ups like Altman did, the cachet you got was worth as much as real earnings. Altman eventually shifted the majority of his money into two other ambitious goals besides AI: extending life and creating limitless energy, betting on two companies. More than $375 million went into Helion and another $180 million into Retro Biosciences, a start-up that was working on adding ten years to the average human lifespan.

If you’re wondering where Altman got the money for that, remember that he put about $3 million into Cruise well before that start-up sold to General Motors for $1.25 billion, earning him a windfall. His place at the top of YC meant he was better positioned than many other venture capitalists to win jackpots like that, getting an intimate view on hundreds of companies who’d already been carefully screened, and in the middle of one of the greatest bull market runs in history. Getting pitched by all those start-ups also helped him see into the future.

A year into heading YC, Altman had all but sealed his reputation as Silicon Valley's new maharishi. He was getting four hundred meeting requests a week. He'd become a magnet for investors and start-up founders who wanted to use him to access other start-ups and partners in Y Combinator or just to meet the even-more-ambitious, sci-fi version of Paul Graham. On blog.samaltman.com, Altman was pontificating on topics that were clearly beyond his area of expertise. He wrote about UFOs and regulation, and he gave advice on how to be a good conversationalist at a dinner party. Don't ask people what they do, Altman wrote. Instead, ask what someone is interested in.

Graham had held weekly "office hours" with YC founders in which he'd talk through some of their problems, dropping pithy words of guidance that tended to follow YC's founding motto: "Make Something People Want." When Altman touched base with start-ups, he steered them toward making things bigger. When the founders of Airbnb, then just a few guys with an app for couch surfers, showed Altman their pitch to investors, Altman told them to take out all the Ms in their presentation and make them Bs, for billions. Either you're ashamed of your pitch or I can't do math, Altman told them with his unblinking stare and large blue eyes.

He advised start-ups to do everything at full throttle and become as driven as he was. "You have to have an almost crazy level of dedication to your company to succeed," he told them. He wrote on his blog that you had to take whatever number measured success and "add a zero." To fix a broken world, founders had to be obsessive about product quality, "relentlessly resourceful," and able to "overcommunicate" with their team. There was no such thing as work-life balance in this world.

Much of what Altman was saying was right. Silicon Valley was where people came to build empires, and you didn't build an empire by working forty hours a week. But his real gift as an entrepreneur was his power to persuade others of his authority. He had drawn the admiration of mentors, from his high school principal to Y Combinator's Graham and Livingston to Peter Thiel, along with thousands of start-up founders. But Altman also had an underlying dissonance: a brilliant mind driven to protect the world that was also emotionally distant from the regular people he sought to save.

This stemmed in part from that sweltering summer's day in 2006, when Altman was lying on the floor in just his gym shorts, freaking out about a deal that wasn't going well. To handle his anxiety, Altman got into meditation, sometimes sitting with his eyes closed and concentrating on just his breath for up to an hour at a time. Over time, he later said, he developed an increasingly diminished sense of self.

"One thing I realized through meditation is that there is no self that I can identify with in any way at all," he told the Art of Accomplishment podcast. "I've heard that of a lot of people spending a lot of time thinking about [powerful AI] get to that in a different way too."

These realizations helped underpin the epiphany he had with his friends while hiking years later, that of course computers would replicate our minds one day. Cognition could happen in a computer, one we'd merge with one day. "Working on AI makes you think about deep philosophical questions in the context of what's going to happen when my mind gets uploaded," he added. "What's going to happen when they're talking to me? Do I want to merge? Do I want to go off exploring the universe? How much of that will still be me?" Altman didn't follow these sci-fi instincts in isolation. He was surrounded by technologists who believed they might also one day upload their consciousness to computer servers, where they could live on in perpetuity.

The idea of death seemed to terrify Altman. He was a self-described prepper and spent a great deal of time and money preparing for a catastrophic global event, like a synthetic virus being released into the world or being attacked by AI. "I try not to think about it too much," he was quoted as telling a group of start-up founders in his *New Yorker* profile. "But I have guns, gold, potassium iodide, antibiotics, batteries, water, gas masks from the Israeli Defense Force, and a big patch of land in Big Sur I can fly to."

He also paid $10,000 to get on the waiting list for Nectome, a Y Combinator start-up that preserved human brains through a high-tech embalming process, so that they could be uploaded to the cloud and then turned into a computer simulation by scientists in the future.

As he invested in more companies pioneering the distant future, Altman seemed to be experiencing what's known as the overview effect, a cognitive shift that astronauts get when they view Earth from space and get an overwhelming sense of awe and self-transcendence. He was increasingly seeing the world as if he was in outer space. Conversations with Altman were marked by deep, searching stares and contemplative pauses, as if he were an observer rather than an engaged participant.

Despite his investments in the future of humankind, he was cultivating a kind of mental and emotional divide between himself and other people. To solve their problems, you needed to be "calm, measured, and pragmatic," he says. Altman often pointed to a short story called "The Gentle Seduction" by the American science fiction author Marc Stiegler, about the future impact of tech on people's lives. The story follows the life of Lisa, a woman who encounters various advancements that "seduce" her into incorporating tech into her daily life.

Toward the end, Lisa and her husband are undergoing the process of uploading their consciousnesses into a computer. It's a risky procedure and the people who merge their minds with this advanced machinery can end up losing themselves, so Lisa weighs the pros and cons, noting that some of her friends who tried it ended up "dying," or getting lost in the digital ether. Stiegler then writes, "Only those who knew caution without fear, only those marked by her elemental form of prudence, made it through."

Altman was struck by that quote and would repeat it to others. The author was saying that to survive the risks of merging with computers, people needed to adopt a mindset that balanced caution and bravery. You were more likely to survive a future threat by being careful and cool-headed, rationally assessing dangers, instead of reacting emotionally and succumbing to panic. The people who thrived in the future would take a detached and informed approach to tech advancements.

Some technologists were leaning too far into anxiety about the future dangers of artificial intelligence, as part of a nascent field of study referred to as "AI safety." While that research was important, some of that panic had turned into fearmongering, and it seemed like these advocates for humanity were letting emotions get the better of them. "Unfortunately, some of the communities involved in AI safety are the people who are least calm," Altman said. "That's a dangerous situation.… It's an extremely high-strung community." But he was also coming to a realization, he says today: "I really wanted to work on AGI." The term AGI had been conjured by Shane Legg a few years earlier, but the idea of creating some kind of cognitive parity between humans and machines had been around for a few decades, partly morphing out of ideas that were first proposed in science fiction. Now it was slowly starting to shift away from being the "crazy talk" that once forced DeepMind's cofounders to discuss their plans in an Italian restaurant, to a serious scientific goal.

The world also needed someone with a more balanced approach to building AI. When Stiegler referred to an "elemental form of prudence," Altman saw his own sensibilities being described—someone who had the wisdom to navigate a complex and potentially dangerous future and who had "caution without fear." He could be the vigilant watchman who stood on the tower's edge, eyes glued on an AI utopia on the horizon and seldom looking at the bustling life just beneath. But he would also become so wrapped up in his mission and self-belief that he wouldn't see the irony of painting himself as cautious, an entrepreneur whose fierce competitive streak would see him hurriedly push artificial intelligence systems out to the public ahead of any other tech company, including Google. Quietly, Altman was also obsessed with being first.

That is why he might never have taken the steps to build that AI utopia if he hadn't been spurred by someone who pioneered the idea. The Silicon Valley entrepreneur needed a rival to spark his own endeavor, and that person was on the other side of the world in England, a brilliant young game designer who was planning to build software so powerful that it could make profound discoveries about science and even God.

CHAPTER 4

A Better Brain

After the collapse of Elixir Studios, Hassabis had become just another failed tech entrepreneur with too-bold dreams. The experience had been painful, but he still had something that he felt was unique to most of the other start-up founders and humans around him: his brain. Hassabis went to great lengths to care for the gray matter in his skull. He played games to exercise it. He avoided alcohol to protect it. He even made his Facebook profile picture an MRI brain scan. Hassabis couldn't help but marvel at its complexity, and in the years after Elixir, he wondered if the brain itself could be the key to making software that was as smart as humans. After all, it was the only proof in the universe that general intelligence was possible, so it made sense to deeply understand it. Was it all physical biology or something more? The answer lay in neuroscience.

Hassabis craved the comfort of certainty, whether that was from the outcome of either winning or losing in games, the moral guidelines of right and wrong that Christianity provided, or the quest for a single framework for the universe that he'd read about in high school. When you could measure something with numbers or rules—that was his sweet spot. "Most of the functions that the brain has, you should be able to mimic in some way with a computer," he would later say in a press interview. "Neuroscience shows you can describe the brain in mechanistic terms." In other words, the frightening complexity of the brain could be boiled down to numbers and data, and be described in the same way as a machine.

To that end, Hassabis took inspiration from Alan Turing, the twentieth-century British computer scientist who came up with the Turing machine. Introduced in 1936, it was essentially a thought experiment, a "machine" that only ever existed in Turing's mind. He envisioned a length of infinite tape that was divided into cells, as well as a tape head that could read and write symbols on the tape, guided by certain rules, until it was told to stop. The idea sounds rudimentary, but as a theory, it was critical in formalizing the concept that computers could use algorithms—or sets of rules—to do things. Given enough time and resources, a Turing machine could be as powerful as any digital computer today. And to Hassabis, it was a perfect proxy for the human mind. "The human brain *is* a Turing machine," he once said.

In 2005, months after closing Elixir, Hassabis jumped into doing a PhD in neuroscience at University College London. His final thesis was relatively short but scientifically exquisite, according to other computer science academics. It was all about memory. Till then, it was thought that the brain's hippocampus mostly processed memories, but Hassabis showed (with the help of other studies of MRI scans in his thesis) that it was also activated during the act of imagination.

In simple terms, this meant that when we had a memory, we were partly imagining it. Our brains weren't just "replaying" past events by retrieving them, as you might take a file out of a filing cabinet, but actively reconstructing them in the way you might paint a picture. The brain was engaged in a much more dynamic and creative process, which went some way to explain why our memories sometimes were plain wrong and could be influenced by our other experiences. Hassabis argued that our brains were using this process of "scene construction" for other types of tasks, like figuring out how to navigate a map or making plans.

His thesis was cited as one of the most important scientific breakthroughs that year by a leading peer-reviewed journal. But Hassabis didn't want to linger in academia. Scholars eager to make Nobel-worthy discoveries spent more than half their time writing grant proposals, and even if they were lucky to get funding for a certain project, most universities didn't have much computing power. To do machine-learning research that was cutting edge, you needed access to some of the world's most powerful computers. Most of them, along with the world's top talent, could only be found at large tech companies. If Hassabis was going to bring together large amounts of brain power to build a kind of modern-day Manhattan Project, he would need to start a business.

The first blueprints came together through conversations over lunch with two other people: Shane Legg and Mustafa Suleyman. Legg was that rare AI enthusiast whose ideas about the future of artificial intelligence almost made those of Hassabis look puny by comparison. He had written a PhD thesis on "machine superintelligence" and his supervisor recommended that he speak to Hassabis afterward.

"I found a kindred spirit," Hassabis remembers. "Shane was someone who had independently arrived at that conclusion that this would be one of the most important things to do ever."

Legg's ideas had already been making waves in the close-knit "singularity" community. These were researchers who believed in a theoretical point in the future when the growth of technology became so advanced that it was unstoppable and uncontrollable. The clearest sign would be when computers became smarter than humans, and Legg believed that would happen sometime around 2030.

His life in frontier science had an unlikely start. Brought up in New Zealand, his parents took him to see an educational psychologist when he was nine after he started struggling at school. The psychologist gave Legg an intelligence test and told his parents, with some annoyance, that he had dyslexia and that his intelligence was off the charts. Once he learned to use a keyboard, Legg zoomed up the school rankings to become one of its highest achievers in math and computer programming.

Tall, slightly stooped, and with closely cropped hair, Legg was twenty-seven when he walked into a bookshop and spotted *The Age of Spiritual Machines* by Ray Kurzweil, which predicted that computers would one day develop free will and have emotional and spiritual experiences.

He read it cover to cover and couldn't stop thinking about Kurzweil's reasoning or his forecast for powerful AI to come about in the late 2020s. Computing power and data were growing exponentially. So long as that kept happening, computers would eventually surpass humans. This correlated with a fundamental principle that underpinned the technology industry itself, known as Moore's Law. It stated that the number of transistors on a microchip would double every two years, an estimate that has held true for the last fifty years.

In the year 2000, when Legg read Kurzweil's book, the dust was still settling from the dotcom bust, so it was hard to believe that computers would continue doubling their capability. But Legg believed the internet would keep growing.

"It was clear that various sensors were going to reduce costs, so there were going to be more and more data that you could then potentially train models on," he says today.

Add all that power and data, and you could train machines to become smarter and smarter. Legg went off to study for a PhD in AI and build up a network of contacts in the field. At one point Ben Goertzel, a singularity believer and AI scientist with long hippy hair, emailed Legg and several other scientists seeking ideas for a book title. It needed to describe artificial intelligence with human capabilities. Legg emailed him back, suggesting a phrase that would become a focal point for Hassabis and, eventually, a handful of the world's largest tech companies: "artificial general intelligence."

For years, people like Hassabis, Legg, and other scientists exploring AI had used terms like *strong AI* or *proper AI* to refer to future software that displayed the same kind of intelligence as humans. But using the word *general* drove home an important point: the human brain was special because of all the different things it could do, from calculating numbers to peeling an orange to writing a poem. Machines could be programmed to do each of those things fairly well, but none could do all of them at once. If a computer could not only crunch numbers but also make predictions, recognize images, talk, generate text, plan, and "imagine," then it might come close to being like a human.

Most AI scientists at the time dismissed the notion that AI could ever reach human parity. Partly this was down to their personal experiences of the hype and failures in AI's history. People would get excited about the possibilities of AI and then find themselves disappointed. A series of booms and busts in AI's history were known as "winters," fallow periods where researchers watched their funding drop as they made achingly slow technological progress. In the 1990s and early 2000s, researchers managed to apply machine learning techniques to narrow tasks like recognizing faces or language, but by the time Hassabis was finishing his PhD in 2009, hardly anyone believed that machines could have *general* intelligence. They'd be laughed out of the room. It was a fringe theory.

Fortunately, Goertzel was on the fringe, and while "artificial general intelligence," or AGI, wasn't snappy, he liked it enough that he slapped the term on his book and helped turn it into a common expression that would go on to help fuel hype about the field.

Language and terminology would end up playing an enormous role in the development of AI, driving interest to sometimes maddening effect. The term itself, *artificial intelligence*, was coined back in 1956 at a workshop at Dartmouth College that was aimed at pulling together ideas about "thinking machines." There were various other names for the new field at the time, like *cybernetics* and *complex information processing*, but *artificial intelligence* stuck. It went on to become one of the most successful marketing terms of all time and spawned a collection of other terms that anthropomorphize machines in our collective consciousness, often lending them more capability than they deserve. It isn't technically accurate, for instance, to suggest that computers can "think" or "learn," but phrases like *neural network*, *deep learning*, and *training* help promote that idea in our minds by lending software humanlike qualities, even when they're only *loosely* inspired by the human brain. The one thing everyone could agree on about Legg's new term, AGI, was that it didn't exist yet.

One other person who believed it could was Mustafa Suleyman. At age twenty-five, the Oxford University dropout was looking for a way to use technology to change the world. He had a brain like a vice, but his areas of expertise were more in policy and philosophy than computer science. Born to a Syrian father and English mother, Suleyman had an overwhelming drive to solve problems. His drive was not to solve minor issues like fixing a broken-down car or rehabilitating someone's knee but large-scale issues that affected all of humanity like poverty or the climate crisis.

Having already cofounded a conflict-resolution firm, he was now interested in studying neuroscience, and Hassabis invited him to some of the informational lunches at University College London. Suleyman already knew Hassabis well. Having grown up in North London, he was a friend of Hassabis's brother, George, and had been a frequent visitor to their home in his teens. The trio had even traveled to Las Vegas to play at a poker tournament in their twenties, coaching one another and splitting the winnings.

When he met Hassabis again, he was struck by his friend's ideas about building a powerful AI system to solve problems, and in Legg's belief in a general intelligence that could tackle almost any issue. Suleyman became excited about what that could mean for societal problems.

The three men would meet at Carluccio's, an Italian chain restaurant near the university, mainly for privacy. "We didn't want people to hear our crazy talk about starting AGI," says Legg.

After some convincing from Hassabis, Legg agreed they probably couldn't build AGI in academia. "We'd be like, professors in our fifties by the time they'd give us any resources to do the sorts of things we wanted to do," Hassabis says. "Companies was what I knew how to do."

To get the necessary scale and resources, they needed to create a start-up. Suleyman had cofounded a company, which meant he knew a thing or two about running a business, as did Hassabis. In 2010, tech companies like Google and Facebook were having the biggest impacts on society, so it made sense to the three men that a technology company would have the greatest chance of modeling the complexity of the world. They hatched an ambitious plan to form a research company that would figure out how to make the most powerful AI anyone had ever seen and then use it to solve global problems.

They named the company DeepMind, made Hassabis the CEO, immediately hired one of Hassabis's top coders from Elixir, and leased office space in an attic across the road from University College London, where Hassabis had done his PhD. The trio had an energy that came from believing in a shared mission, though they had different motivations. Legg moved in circles where the goal was to merge as many people with AGI as possible, Suleyman wanted to solve societal problems, and Hassabis wanted to go down in history having made fundamental discoveries about the universe.

It wasn't long before there were debates over their different goals. Suleyman was eager for Hassabis to read a book that had shaped his view of the world. Called *The Ingenuity Gap*, it was published in 2000 by Canadian academic Thomas Homer-Dixon and argued that the utter complexity of modern-day problems, from climate change to political instability, was outpacing our ability to come up with solutions. The result was an ingenuity gap, and humans needed to innovate in areas like technology if they wanted to close it. That's where AI could fit in, Suleyman figured.

Hassabis shook his head. "You're missing the bigger picture," he once told him, according to someone who heard about the conversation. Hassabis seemed to believe Suleyman's view on AI was too narrowly focused on the present and that AGI would be better used to help DeepMind understand where humans had come from and what their purpose was. Hassabis suggested that climate change, for instance, was humanity's destiny and that Earth probably couldn't carry everyone on it into a long-term future. He said that trying to solve current problems was like playing around in the margins when such events were probably inevitable. He didn't believe that superintelligent machines would go rogue and kill humans, as some were starting to fear. Instead, AGI would solve some of our most profound problems once he had built it.

Hassabis summed up that view in DeepMind's tagline: "Solve intelligence and use it to solve everything else." He put it on their slide deck for investors.

But Suleyman disagreed with that vision. One day when Hassabis wasn't around, he told one of DeepMind's early staff members to change it on a slide presentation. It now read: "Solve intelligence and use it to make the world a better place."

Hassabis didn't like that either. Later, when Hassabis was in the office again, he asked the same staff member to change it back. Now it said, "Use it to solve everything else” once again. As the two tussled on the company’s mission with their staff as proxies, they were avoiding direct confrontation in the most British way possible.

Suleyman wanted to build AGI in the way Sam Altman eventually would, by sending it out into the world to be immediately useful. It was better to gather feedback from the real world and improve than work in isolation to try to build the perfect system. But Hassabis wanted to run DeepMind with the end in mind, just in the same way he played chess. The prize wasn’t solving only real-world problems but the mysteries that had perplexed humankind for generations. What was our purpose, and did we come from a divine being?

Hassabis is coy when asked if he believes in God. “I do feel there’s mystery in the universe,” he says. “I wouldn’t say it’s like traditional God.” He says that Albert Einstein believed in “the God of Spinoza, and maybe I’d give a similar sort of answer.”

Baruch Spinoza was a seventeenth-century philosopher who proposed that God was effectively nature and everything that existed, rather than a separate being. It was a pantheistic view. “Spinoza thinks of nature as the embodiment of whatever God is,” Hassabis says. “So doing science is exploring that mystery.”

It wasn’t crazy to think that creating AGI could become a spiritual or quasi-religious experience akin to a divine discovery, especially if you took Spinoza’s view that God was equivalent to the laws of nature. By using AI to delve into those laws and understand the universe, you could theoretically puzzle out a designer. With its ability to analyze vast amounts of data, AI could study some of the most complex systems in the universe, from quantum mechanics to cosmic phenomena, and dig out insights into the intricate nature of existence. Using AI to create a simulation that mimicked the complexity of the universe could also reveal parallels to how our universe operated.

And if AGI research led to the conclusion that our universe was a simulation—as Kurzweil himself has proposed—the original programmer could well be a godlike entity. Similarly, if humans created a powerful machine that ingested and analyzed all available information about physics and the universe, that machine could also propose new theories that suggested the existence of a higher power. It might just answer deep existential questions that pointed to a divine entity. There were myriad ways that with greater capabilities and intelligence, AI could unlock one of humanity’s most profound secrets.

Hassabis’s religious background might also have made him more receptive to the idea of an AI oracle. A 2023 study from the University of Virginia that involved more than fifty thousand participants from twenty-one countries found that people who believed in God or thought about God more than others were more likely to trust advice from an AI system like ChatGPT. According to the researchers, these people were more receptive to AI guidance because they tended to have greater feelings of humility. They were also quick to recognize human flaws.

Hassabis would sometimes talk to his early DeepMind colleagues about God as his mind churned with questions about humankind’s origins. Several people who worked with Hassabis or know him personally say he was a devout Christian for years, and one says his primary reason for building AGI was to discover God.

"We had many discussions about God," says a colleague who worked with Hassabis around the time he cofounded DeepMind. "Could we create a machine that could work backwards to make sense of the universe? AGI would give you an insight into where we came from, and what is God." Hassabis also believed he was running a modern-day Manhattan Project. He'd read *The Making of the Atomic Bomb*, and it inspired him to structure DeepMind's team as Robert Oppenheimer had: focusing teams of scientists on subsections of a bigger problem, according to two former DeepMind staffers.

But to make such an ambitious discovery, Hassabis needed money to grow DeepMind. Unfortunately, British investors were only offering paltry sums of £20,000 or £50,000 for equity in his new start-up. That wasn't nearly enough money to hire the talent he needed to build AGI, never mind access the powerful computers he'd need too. It didn't help that his business idea of building the world's most powerful AI system seemed outlandish and overly ambitious in button-down Britain. In the UK, tech start-ups tended to chase "sensible" business ideas that would make money more quickly, like building a financial app for trading stocks and bonds. Hassabis and his cofounders had little choice but to look to Silicon Valley, where investors were willing to bet bigger amounts of money for more futuristic ideas.

Legg fortunately had an in. He'd been invited to speak at the Singularity Summit in June 2010, an annual conference cofounded by Kurzweil, the author who had captivated him as a young man, and Peter Thiel, the billionaire investor who liked putting his money into pioneering new technology. This was a conference where some of AI's most unconventional scientists talked about technology's awe-inspiring power and risks. Thiel set the tone for the event, and he was an idealist. He didn't think the singularity, that moment in the future when AI would irreversibly change humanity, would be a problem—quite the opposite. He worried it would take too long to get here and that the world needed powerful AI to ward off economic decline.

With his deep pockets and enthusiasm for ambitious projects, Thiel was the perfect person to fund DeepMind. "We needed someone crazy enough to fund an AGI company," Legg remembers. "We needed somebody who had the resources to not sweat a few million and liked super ambitious stuff. They also had to be massively contrarian because every professor [Hassabis] talked to would tell him, 'Absolutely do not even think about doing this.'"

Thiel was so contrarian that he was often at odds with the rest of Silicon Valley, which was already full of unconventional thinkers. While most of the region voted liberal, he veered to the right, becoming one of President Donald Trump's top donors. While most entrepreneurs believed competition drove innovation, Thiel argued in his book *Zero to One* that monopolies did that better. He scorned the conventional routes to success, encouraging smart, entrepreneurial kids to drop out of college and join his

Thiel Fellowship. And his wacky pursuits for longevity and the singularity meant he fit the "crazy" criteria the DeepMind founders were after.

The trio decided to pitch to Thiel at the Singularity Summit. He was funding the event, and they figured that meant he'd be sitting in the front row. Legg asked the summit's organizers if he could share his speaking slot with Hassabis. That way Thiel could hear directly from the former chess champ about building AGI with the human brain as inspiration.

Dressed in a wine-red sweater and black slacks, Hassabis was shaking when he took to the stage at the summit in a San Francisco hotel for a moment that would decide if his new company lived or died. But when he looked out at the crowd of hundreds of people, Thiel wasn't in the front row. He wasn't in the audience at all.

The founders thought they'd blown their chance, but then Legg got an exclusive invite to a party at Thiel's Bay Area mansion, and he managed to get his cofounders an invite. Hassabis had learned that Thiel liked chess. Thiel had at one time been one of the best under-thirteen chess players in the United States. Now there was some common ground and an opportunity to spark some intrigue. During the party, Hassabis struck up a conversation with Thiel and casually mentioned the game, according to an account he's shared several times with the press.

"I think one reason why chess has survived so successfully over generations is because the knight and bishop are perfectly balanced," Hassabis told Thiel as canapés were being passed around. "I think that causes all the creative asymmetric tension."

Thiel's interest was piqued. "Why don't you come back tomorrow and do a proper pitch?" he said. The trip turned out to be a success. Thiel invested £1.4 million into helping DeepMind bring on the singularity.

As Hassabis tried to raise more money to grow his AI company, he was faced with an awkward situation for an entrepreneur. His first investors weren't necessarily backing him because they wanted to make money but because they had an almost moral belief about artificial intelligence. It meant that he'd be facing a messier kind of pressure over how he ran his company, to not just make money but also develop AI in a way that conformed to various dogmas.

One belief system gathering steam at the time was that AI needed to be built with great caution so that it wouldn't break away from human control and try to destroy its creators. Those were the worries of another wealthy donor—with opposing views to Thiel—who also wanted to back DeepMind. Hassabis met that donor when he went to Oxford for Winter Intelligence, a conference that was on the fringe of computer science research, where some of the field's most radical thinkers were giving talks about the challenges of controlling superintelligent AI. Moments after Hassabis gave his talk, a man with short blond hair and a Nordic-sounding accent approached him.

"Hi," the man said as he approached Hassabis and held out his hand. "I'm Jaan. I'm the [cofounder] of Skype."

Originally from Estonia, Jaan Tallinn was a computer programmer who developed the peer-to-peer technology underpinning Kazaa, one of the first file-sharing services used to pirate music and movies in the early 2000s. He repurposed that technology for Skype and took a stake in the free calling service before getting a massive windfall when eBay bought Skype for $2.5 billion in 2005. Now he was sprinkling some of his winnings into other start-ups. When Tallinn heard Hassabis speak, his ears pricked up. He'd recently become passionate about the dangers of powerful artificial intelligence.

Tallinn caught the AI bug two years earlier in the spring of 2009, when he'd been reading some essays on a website called LessWrong. The online forum was a tight-knit community of members, many of them software engineers, who worried that AI posed an existential risk to humanity. Their guru and the site's founder was a bearded libertarian named Eliezer Yudkowsky, a high-functioning high school dropout who taught himself the fundamentals of artificial intelligence research and philosophy and whose essays enthralled the site's members. Yudkowsky was the kind of person Altman had been referring to when he called the AI safety community "high strung." He believed that AI was more likely than anyone realized to annihilate humanity.

Once it got to a certain level of intelligence, for instance, AI could strategically hide its capabilities until it was too late for humans to control its actions. It could then manipulate financial markets, take control of communications networks, or disable critical infrastructure like electrical grids. The people who were building AI often had no idea that they were bringing the world closer and closer to its destruction, Yudkowsky wrote.

Tallinn found himself perturbed by some of these essays. He'd already been mulling over the conclusions of a book he had just read by Roger Penrose called *Shadows of the Mind*. In it, the renowned physicist and mathematician argued that the human mind could perform tasks that no computer ever could. The ideas that Hassabis and others had proposed about the brain being "mechanistic" and a useful inspiration for building AI didn't hold water, because the human brain was unique. It was virtually impossible to replicate.

But something was nagging at Tallinn about that conclusion. What if you *could* simulate the human mind as an artificial intelligence? Wouldn't that mean we were building something potentially dangerous? The Skype founder wanted to hear more from Yudkowsky, so he jotted down a list of questions, trying to pick holes in some of the doom-laden arguments. The best way to figure out if any of it was true was to meet the LessWrong founder himself.

Luckily, Tallinn was planning to fly to San Francisco for a meeting, so he sent Yudkowsky an email, asking if he'd like to get together for a chat. The American replied and agreed to coffee. When they sat down at a café in the city of Millbrae, a short drive from San Francisco International Airport, Tallinn started going through his questions. If AI was potentially dangerous, why couldn't we just build it on virtual machines to separate it from other computer systems? Surely that would stop the AI from infiltrating our physical infrastructure and shutting down an electrical grid or manipulating financial markets.

Yudkowsky immediately had an answer. “It wouldn’t really be virtual,” he replied, sipping his drink. Electrons could flow in all sorts of different directions, which meant there was always going to be a way for powerful AI systems to touch and change the configuration of hardware.

This confirmed what Tallinn was worrying about. One day, he thought, AI could develop its own infrastructure and its own computer substrate. The possibilities after that were terrifying in their scope.

“It could terraform and geo-engineer the planet and possibly the sun,” he says today. When scientists argued that AI was just math and that there was no need to fear it, Tallinn liked to point to the analogy of the tiger. “You could argue that a tiger is just a bunch of biochemical reactions, and there’s no point in being afraid of those.” But a tiger is also a collection of atoms and cells that can do plenty of damage if not kept in check. Similarly, AI might just be a collection of advanced math and computer code, but when put together in the wrong way, it could be incredibly dangerous.

By the time Tallinn found himself listening to Hassabis speak at the Oxford conference two years later, he’d become a convert to the teachings of AI doom. He’d been devouring Yudkowsky’s essays since that meeting in the café and had steeped himself in a new field of research called AI alignment, where scientists and philosophers were figuring out how best to “align” artificial intelligence systems with human goals.

“I had been alignment pilled,” Tallinn remembers. And he now believed in some of the more extreme scenarios about a future AI that Yudkowsky had been sketching out.

After some initial chitchat, Tallinn wanted to see if Hassabis would be willing to work more closely together. “Do you want to have a Skype meeting sometime?” he asked the British entrepreneur.

Hassabis and the wealthy Estonian spoke again, and Tallinn eventually became one of DeepMind’s first investors alongside Peter Thiel. His goal wasn’t just to make money but to keep an eye on Hassabis’s progress and make sure he didn’t inadvertently create a horrifying, rogue AI. Tallinn saw himself as an evangelist for Yudkowsky’s ideas. He wanted to use his credibility as a deep-pocketed investor to help expose his warnings to the world’s most promising AI builders.

“Eliezer is an autodidact and didn’t have much clout outside his small community,” Tallinn explains. “I thought I could start selling those arguments to people who wouldn’t listen to Eliezer but would listen to me.”

Once he was an investor, Tallinn pushed DeepMind to focus on safety. He knew that Hassabis wasn’t as worried about the apocalyptic risks of AI as he was, so he put pressure on the company to hire a team of people that would study all the different ways they could design AI to keep it aligned with human values and prevent it from going off the rails.

DeepMind was about to get another investor with even deeper pockets who also wanted to steer it in a safe direction. Back in Silicon Valley, rumors were swirling of Peter Thiel's involvement in a promising but secretive new start-up in London, UK, that was trying to build artificial general intelligence. Some of the region's other technology billionaires were starting to hear about it, and one of them was Elon Musk. In 2012, two years after cofounding DeepMind, Hassabis was mingling at an exclusive conference in California that Thiel had organized when he bumped into Musk.

"We hit it off straight away," Hassabis says. The British entrepreneur knew that this could be an opportunity to raise more money to expand DeepMind's research—and he also really wanted to see Musk's rocket factory. Musk was establishing himself as a maverick tycoon who wanted to send humans to Mars with his company SpaceX. Hassabis arranged to meet Musk at the company's headquarters in Los Angeles.

Later the two men were sitting across from each other at the company's canteen, amid rocket parts, and found themselves having a debate over who worked on the most historically important project: interplanetary colonization or developing super AI.

"Humans will need to be able to escape to Mars if AI gets out of control," Musk said, according to a *Vanity Fair* article that recounted the meeting.

"I think the AI would be able to follow everybody to Mars," replied Hassabis, who seemed amused. Musk was not. While Tallinn had been influenced by the online writings of Yudkowsky, Musk had been moved by someone else: an Oxford philosopher named Nick Bostrom.

Bostrom had written a book called *Superintelligence*, and it was causing a stir among people working on AI and frontier technology. In the book, Bostrom warned that building "general" or powerful AI could lead to a disastrous outcome for humans, but he pointed out that it might not necessarily destroy us because it was malevolent or power-hungry. It might just be trying to do its job. For instance, if it was given the task of making as many paper clips as possible, it might decide to convert all of Earth's resources and even humans into paper clips as the most effective way to fulfill its objective. His anecdote spawned a saying in AI circles, that we need to avoid becoming "paper-clipped."

Musk went ahead and put some money into DeepMind too. While Hassabis finally had some financial security, it wasn't a lot. He was still pursuing something that was highly experimental and so crazy that even some of the world's richest men didn't want to bet too much money on his success. Their money also came with ideological strings attached: Tallinn and Musk watched DeepMind's work with an unusual amount of suspicion and wariness for investors. They wanted DeepMind to financially succeed, sure, but they also didn't want DeepMind to build too quickly or in a way that would put humanity in danger. That put Hassabis in an awkward position. He was grateful for their money, but he didn't believe in the doom-heavy scenarios that Tallinn and Musk did.

As it happened, that sense of financial security didn't last very long. Hassabis and Suleyman were struggling to make enough money to cover the costs of paying the best AI minds in the world, and some of their ideas for generating revenue were all over the place. They tried setting up a website that used deep learning—a type of machine learning that DeepMind initially specialized in—to give people fashion advice and recommend clothes. Then Hassabis asked some of his staff who he'd managed at Elixir, and who now worked at DeepMind, to design a video game. The engineers pulled together a space-faring adventure where a crew of astronauts had to race to the moon in a rocket, according to a former DeepMind staffer. They were just gearing up to release the game as an iPhone app when Hassabis was presented with a new opportunity, something that would give him the financial backing he needed to make AGI a reality. It was an offer from Facebook.

Mark Zuckerberg was on an acquisition tear. About a year earlier, he had bought Instagram for $1 billion in what would become a masterstroke of social media consolidation. And he was just months away from paying an eye-watering $19 billion to the founders of WhatsApp. He was ready to spend whatever he needed to grow the Facebook empire, and artificial intelligence was going to be an important part of that. Facebook made about 98 percent of its money from selling ads, but to sell more advertisements and keep growing, Zuckerberg needed people to spend more and more time on his sites. DeepMind's dozens of talented AI scientists could help. With smarter recommendation systems that could trawl through the personal data of their users, smarter algorithms behind Facebook and Instagram could show people the right pictures, posts, and videos to keep them scrolling for longer.

Zuckerberg offered Hassabis $800 million for DeepMind, not including the bonus that start-up founders typically got after staying with their acquired company for four or five years, according to a person familiar with the offer. It was a generous bid and more money than Hassabis had ever dreamed of. He now found himself at a crossroads. Until now, DeepMind's money had been coming from people who wanted him to build AI as carefully as possible. Now it could come from someone who wanted them to build it much more quickly. Facebook's motto after all was "Move fast and break things."

Hassabis and Suleyman talked about how to approach the situation. AGI was going to be more powerful than even Zuckerberg realized, and they felt they needed something in place to prevent a big corporate acquirer from steering AI in a potentially harmful direction. They couldn't just have Facebook sign a contract and promise not to misuse AGI. Thinking back to his previous work with nonprofit groups, Suleyman told Hassabis and Legg that they needed some sort of governance structure that could keep a close eye on Facebook and make sure it was careful in how it used DeepMind's technology.

Public companies typically have a board of directors whose job is to represent the interests of shareholders. Those directors will meet every quarter and scrutinize the company's actions to make sure that it's doing all the right things to help its stock go up instead of down. Suleyman told his cofounders that DeepMind should have a different kind of board to deal with technology as transformative as AI. Instead of focusing on money, their job would be to make sure DeepMind was building AI as safely and ethically as possible. Hassabis and Legg weren't convinced at first, but Suleyman was persuasive and they eventually agreed to the idea.

Hassabis went back to Zuckerberg and told him that if they were going to sell, DeepMind would need to have this ethics and safety board in place and that it would need to have separate legal authority to control any superintelligent AI that DeepMind eventually built. Zuckerberg balked at the demand. He wanted to grow Facebook's advertising business and "connect the world" through his various social media platforms, not run a separate AI company with a bunch of ethical protocols and its own grand mission. The talks fell apart.

Outwardly, Hassabis told his employees that DeepMind would stay independent for another twenty years. But privately he was tired of fundraising and frustrated that he was only spending a fraction of his time on actual research. Having just rejected a huge offer from Zuckerberg, it was hard to ignore how much money he could make from selling to a company in Silicon Valley, especially now that Big Tech was suddenly salivating over AI. Senior executives from Silicon Valley's biggest companies, including a billionaire or two, were now calling DeepMind's researchers regularly to try to poach them. Many of the company's staff were experts in deep learning, which for years had seemed like a backwater in the field until just recently.

The turning point had come in 2012. A Stanford AI professor named Fei-Fei Li had created an annual challenge for academics called ImageNet, to which researchers submitted AI models that tried to visually recognize images of cats, furniture, cars, and more. That year, scientist Geoffrey Hinton's team of researchers used deep learning to create a model that was far more accurate than anything before, and their results stunned the AI field. Suddenly everybody wanted to hire experts in this deep-learning AI theory inspired by how the brain recognized patterns.

It was a tiny field with just a few dozen experts, says Legg. "We'd hired quite a few of them." Hassabis was paying them about $100,000 a year, but tech giants like Google and Facebook would pay several times that much. "We were having seriously famous people cold-calling our researchers offering them three times their salary," Legg remembers. Zuckerberg was one of those famous people, according to a former DeepMind employee. "We had to sell otherwise we would have been torn to pieces." And with Hassabis eager to be the first to build AGI, he couldn't wait for the better-resourced tech companies to get there before him.

Out of nowhere came another offer to buy DeepMind, this time from its investor Elon Musk. The billionaire offered to pay for the company with shares of Tesla, the electric car company he'd been running for the last five years, according to a person who was familiar with the deal. Musk had been a hands-off investor and only occasionally checked in with Hassabis. Despite his growing fears about the dangers of AI, the billionaire’s commercial goals were also at the forefront of his mind. He wanted Tesla cars to be the first in the world to successfully use self-driving technology, which meant he needed more cutting-edge experts in artificial intelligence. Now he could get an elite army of them by purchasing DeepMind.

But once again DeepMind’s founders were wary. Getting paid in Tesla stock didn’t seem all that appealing. They also felt uneasy about someone like Musk taking control of AGI. Although he was just starting to gain mainstream fame as a forward-thinking tycoon, Musk had a reputation in tech circles for being capricious, firing staffers out of the blue and ousting the cofounder of Tesla.

As much as the DeepMind founders appreciated his investment and connections, they were leery of his erratic behavior. They declined his offer, too, not realizing how much the thin-skinned Musk didn’t like it when people said no or how much that decision might come back to haunt them. Soon enough though, Hassabis got another email. It was from Google.

CHAPTER 5

For Utopia, for Money

The message came from an executive at Google's headquarters more than five thousand miles away in sunny Mountain View, California. As Hassabis opened it up on his computer in London, he saw an invitation to meet with Larry Page, Google's CEO. Page had cofounded Google with a fellow Stanford University PhD student named Sergey Brin in 1998. The two had wanted to improve the way people searched the internet, and they did that by making an algorithm called PageRank, which classified web pages based on their relevance and interconnections. Starting in a friend's garage in Menlo Park, California, they eventually created one of the world's biggest technology companies.

But when it came to how Google made money today, that process wasn't very high-tech or innovative: it had become an enormous advertising company, like Facebook. The vast majority of Google's profits and revenues came from tracking people's personal information to target them with ads, through search, YouTube, and Gmail, and on millions of websites and apps that used the Google Display Network.

There was something a little disconcerting about that for someone like Hassabis, who wanted to use AI to help the world. But he also knew that if he didn't bite, Google could end up poaching his staff and maybe build AGI without him. It already had hundreds more engineers working on artificial intelligence, and Hassabis decided that he couldn't decline the meeting request from California.

When he met Page, Hassabis felt like he was speaking to another kindred spirit. Before him was an introverted mathematics grad with dark, bushy eyebrows, who dressed in casual shirts and shorts. Throughout his time building Google, Page had nurtured a dream to create powerful artificial intelligence too. "He told me that he always thought of Google as an AI company, even when he was in that garage in 1998," Hassabis remembers.

Partly it was personal, as Page's father had been a professor of artificial intelligence and computer science until his death in 1996. That made him a kind of second-generation AI technologist. Page admired how serious Hassabis was about building AGI, and he didn't think it was a wacko idea. He'd already green-lit another internal effort at Google to build humanlike AI, too, an effort that would end up sparking an intense rivalry with Hassabis down the line.

Page's project, which Hassabis didn't know about at the time, was called Google Brain. It had come about as a proposal from Andrew Ng, a gentle-voiced Stanford University professor who wanted to build more advanced AI systems from inside Google. In 2011, a few years before Google reached out to DeepMind, this professor had sent a four-page document to Page titled "Neuroscience-Informed Deep Learning." Professor Ng was hoping the Google CEO would greenlight a project for him to build "general purpose" AI systems, which was what Hassabis was working on in England.

It turned out that Ng and Hassabis approached their goals with similar methods, both looking to neuroscience as an inspiration for building AGI. In his proposal, the Stanford professor told Page that he would build "increasingly accurate approximations to small parts of the mammalian brain."

Even for someone like Ng, who was already a leading figure in the field of artificial intelligence working for one of the world's most prestigious universities, the idea of building AGI at the time was controversial. "My friends were advising me that this was just kind of weird. They said, 'It's not good for your career,'" Ng remembers.

In one sense, they were right. When it came to the science, there were some problems with Ng's and Hassabis's obsession with the human brain. In theory, it made sense to use our gray matter as a template for artificial intelligence, but copying what we find in biology doesn't always work. Think of those very first attempts at creating flying machines and the inventors who built contraptions that mimicked the mechanics of birds. They ended up flapping their bulky wing machines straight into the ground. Other computer scientists had been hitting a wall in their efforts to copy the brain too closely. In 2013, neuroscientist Henry Markham said in a TED Talk that he'd figured out how to simulate an entire human brain on supercomputers and that he would achieve it within a decade. Ten years later, his Human Brain Project had cost more than $1 billion and largely failed.

Over the years, Ng, Hassabis, and other AI scientists would realize how difficult it was to emulate the brain when our understanding of it was still so incomplete, from the functions of neutrons to the dynamics of brain regions. Though we knew there were around ninety billion neurons constantly firing in our skulls, we still didn't know how that information was being processed.

"In hindsight, trying to be so true to biology was a mistake," Ng says. But Ng's research got another part of the science very right: making his neural networks bigger.

A neural network is a type of software that gets built by being trained over and over with lots of data. Once it's been trained, it can recognize faces, predict chess moves, or recommend your next Netflix movie. Also known as a "model," a neural network is often made up of many different layers and nodes that process information in a vaguely similar way to our brain's neurons. The more the model is trained, the better those nodes get at predicting or recognizing things.

Ng discovered that these models could do more things if they had more nodes, layers, and data to train on. Years later, OpenAI would make a similar discovery about the importance of "scaling up" on these key ingredients. During his experiments at Stanford, Ng noticed that his deep-learning models performed much better when they were bigger. The results excited him, and then compelled him to send his four-page proposal to Page, suggesting he might be able to build "large-scale brain simulations" as a step toward "human-level AI."

Page loved the idea and approved it, bringing Ng on board to lead Google's most cutting-edge AI research project yet. But a few years later, Google Brain didn't look like it was on course to build AGI. Instead, it was helping Google improve its targeted advertising business—making its ads even more creepily accurate for users by getting better at predicting what people would want to click on—and growing the company's revenues. Ng admits that wasn't what he'd been aiming for when he sent his proposal to Page. "It is not the most inspiring thing I've worked on," he says.

What Ng had really wanted to do with his scientific research was free humanity from mental drudgery, in the same way the Industrial Revolution had liberated us from constant physical labor. Stronger AI systems would do the same for professional workers, he believed, "so we can all pursue intellectually more exciting, high-level tasks."

But Ng's approach to doing that was where he differed from Hassabis. While the British entrepreneur wanted as much independence from the advertising giant as possible, Professor Ng was happy to work inside the belly of the beast at Google. In that sense, Ng had done Hassabis a huge favor. By basing himself in the Google mothership, Ng's research was already on track to contribute to the company's ad business so that DeepMind didn't immediately have to.

By the time Google first reached out to DeepMind about buying the company in late 2013, Ng's researchers were already sucked into building sophisticated AI models to power Google's advertising tools, drawing them away from Ng's loftier goal of building all-powerful AI that could free humanity from drudgery. Now as Page flew to London to negotiate the purchase of DeepMind, he knew he could spend some of Google's money on something a little more out there.

The DeepMind founders greeted the Google billionaire at their London office, giving a presentation about the company's research so far, according to *Genius Makers*, the book by *New York Times* writer Cade Metz. Hassabis described how his team had developed a new technique called reinforcement learning to train an AI system to master the retro Atari game *Breakout*. In the game, you hit a ball up into a wall of bricks with a paddle that you slid from side to side. Within about two hours, the system had learned to knock the ball into exactly the right spot so that it would break a tunnel into the narrow space behind the top row of bricks, knocking out handfuls of them at a time. Page was impressed.

As a technique, reinforcement learning wasn't all that different to how you might reward a dog with treats whenever it sits on command. In training AI, you would similarly reward the model, perhaps a numerical signal like a +1, to show that a certain outcome was good. Through repeated trial and error, and playing hundreds of games over and over, the system learned what worked and what didn't. It was an elegantly simple idea wrapped in highly sophisticated computer code.

Legg then gave a presentation to Page about where that could lead next: applying these techniques to the real world. In the same way their system had mastered a video game, they could similarly teach a robot to find its way around a home or an autonomous agent to navigate the English language. That's where DeepMind's discoveries and AGI itself would eventually have the most impact. Page and his team were sold.

Page led the deal negotiations with Hassabis and his cofounders, knowing that they'd already rejected a big offer from Facebook. He was about to find out why. Hassabis said he had two big conditions for selling. First, he and his cofounders didn't want Google to ever use DeepMind's technology for military purposes, whether that was for steering autonomous drones or weapons or supporting soldiers in the field. He and his cofounders saw these as ethical red lines that Google should never cross.

Second, they wanted Google's leaders to sign what they called an ethics and safety agreement. Drawn up by lawyers in London, it was a contract that gave control of any future artificial general intelligence technology that DeepMind created to an ethics board that Hassabis and his cofounder Suleyman would pull together. They still only had a vague idea of who should be on it, but they wanted it to have complete legal oversight of the powerful AI they would eventually build.

"If we were successful, [AGI] would need careful handling," Hassabis says about the board that he and his cofounders wanted. "Because it's such an all-purpose technology, this could be one of the most powerful ever, and we wanted to make sure we were aligned with people who were also going to take the responsibility of that seriously."

Not surprisingly, it took months of difficult negotiations to get Google to agree to the very same condition that had been a deal-breaker for Facebook. Buying DeepMind meant Page could own the first company to build AGI. He knew that if this ethics board had legal control of that technology, it could become much harder for Google as a company to profit from it, but in the end, Page's idealistic outlook won out. They would find a way to make it work. He agreed to DeepMind's demands for an ethics board as part of the acquisition.

AGI didn't only need careful handling because of where a corporate giant could take it in the future. It was also at the center of several growing ideologies that could steer the technology in different directions. Hassabis had gotten a taste of that from investors like Peter Thiel, who wanted AI development to move faster, and Jaan Tallinn, who feared the young British entrepreneur might spark an apocalypse.

AI's mind-bending potential gave it an almost religious attraction to people with strong beliefs about how it should be used. Over the next few years, these ideological forces would collide with the innovators and the corporate monopolies who were battling to control AGI, becoming an unpredictable hazard for the technology. They would push Sam Altman out of OpenAI, for instance, and paradoxically boost the commercial efforts of companies, painting an apocalyptic picture of AI's power that ended up making the software more attractive to businesses. Mixed in with the world of business and profit, more AI builders were finding themselves devoutly following different dogmas, from building AI as quickly as possible to bring about utopia to stoking fears that it could cause Armageddon.

As a strategic thinker who liked to hedge his bets, Hassabis found himself largely outside of these warring dogmas, thanks in part to his own unique goals of making big and possibly divine discoveries with AGI, according to people who knew him. Suleyman was also more concerned about societal problems that AI could cause sooner. Of the three cofounders, Shane Legg was most aligned with the more extreme ideologies linked to the pursuit of AGI, including one that had been decades in the making, according to his former colleagues. Known as transhumanism, the idea had controversial roots and a history that helped explain why AI's builders sometimes neglected the nasty, more current side effects of the technology.

The basic premise of transhumanism is that the human race is currently second-rate. With the right scientific discoveries and technology, we might one day evolve beyond our physical and mental limits into a new, more intelligent species. We'll be smarter and more creative, and we'll live longer. We might even manage to meld our minds with computers and explore the galaxy.

The core idea stems back to the 1940s and 1960s when an evolutionary biologist named Julian Huxley joined and ran the British Eugenics Society. The eugenics movement proposed that humans should improve themselves through selective breeding, and it flourished in British universities and among the country's intellectual and upper classes. Huxley himself came from an aristocratic family (his brother Aldous wrote *Brave New World*), and he believed society's upper crust was genetically superior. Lower-class people needed to be weeded out like a bad crop and subjected to forced sterilization. "[They] are reproducing too fast," Huxley wrote.

When the Nazis latched on to the eugenics movement, Huxley decided it needed a rebrand. He coined a new term, *transhumanism*, in an essay saying that alongside proper breeding, humanity could also "transcend itself" through science and technology. The movement picked up steam in the 1980s and 1990s when the growing field of artificial intelligence offered a tantalizing new possibility: perhaps scientists could enhance the human mind by merging it with intelligent machines.

This idea was crystallized in the concept of the singularity, a point in the future when AI and technology became so advanced that humankind would undergo dramatic and irreversible change, merging with machines and enhancing themselves with technology. The idea captivated Legg through the book that he read when he was younger as well as DeepMind's wealthy backer, Peter Thiel. So keen were technologists to experience this utopia that some, like Altman and Thiel, had signed up with different companies to have their brain or whole body cryopreserved in case they couldn't manage that mind-meld before death. "I don't necessarily expect it to work," Thiel told journalist Bari Weiss on her podcast. "But I think it's the sort of thing we're supposed to try to do."

The problem with some of these ideas was that, over the years, their followers grew increasingly zealous. Some so-called AI accelerationists, for instance, believe that scientists have a moral imperative to work as quickly as possible to build AGI to create a posthuman paradise, a kind of rapture for nerds. If it was built in their lifetimes, they could live forever. But speeding up AI's development could also mean cutting corners and making technology that harmed certain groups of people or that could spin out of control.

That's where others took the opposite stance, believing that AI represented a kind of devil figure of the future that needed to be stopped. Eliezer Yudkowsky, the bearded libertarian who helped radicalize Jaan Tallinn over coffee, was a leading figure in that ideological movement, which he gave increasing momentum through his site LessWrong. By the time Google had bought DeepMind in 2014, hundreds of people, including AI researchers, were engaging in philosophical debates on the site over how they could prevent a powerful superintelligence of the future from causing annihilation. LessWrong had become the internet's most influential hub for AI apocalypse fears, and some press reports pointed out that it had all the trappings of a modern doomsday cult. When a member suggested a novel way AI could destroy humans in the future, Yudkowsky publicly laid into them in all caps and kicked them out of the group.

Over time, so-called AI doomers gained enough support among wealthy technologists to pour money into starting companies and shaping government policy to help their agenda. And Yudkowsky's site would become so influential that many of its avid readers would end up joining OpenAI.

But perhaps the most disturbing ideologies that were starting to percolate around AGI were those focused on creating a near-perfect human species in digital form. This idea was popularized in part by Bostrom's *Superintelligence*. The book had a paradoxical impact on the AI field. It managed to stoke greater fear about the destruction that AI could bring by "paper-clipping us," but it also predicted a glorious utopia that powerful AI could usher in if created properly. One of the most captivating features of that utopia, according to Bostrom, was "posthumans" who would have "vastly greater capacities than present human beings" and exist in digital substrates. In this digital utopia, humans could experience environments that defied the laws of physics, like dying unaided or exploring fantastical worlds. They could choose to relive cherished memories, create new adventures, or even experience different forms of consciousness. Interactions with other humans would become more profound, because these new humans would be able to share thoughts and emotions with one another directly, leading to deeper connections.

These ideas were irresistible to some people in Silicon Valley, who believed such fantastical ways of life were achievable with the right algorithms. By painting a future that could look like either heaven or hell, Bostrom sparked a prevailing wisdom that would eventually drive the Silicon Valley AI builders like Sam Altman to race to build AGI before Demis Hassabis did in London: *they* had to build AGI first because only *they* could do so safely. If not, someone else might build AGI that was misaligned with human values and annihilate not just the few billion people living on Earth but potentially trillions of perfect new digital human beings in the future. We would all lose the opportunity to live in nirvana. Along the way, Bostrom's ideas would also have dangerous repercussions as they drew attention away from studying how artificial intelligence could harm people living in the present.

As these modern-day technological ideologies coincided with DeepMind's negotiations with Google, a hard truth was coming to bear. Figuring out a responsible form of stewardship for AI was becoming fraught for tech companies. Different objectives were on track to crash into one another, driven by an almost religious zealotry on one side and an unstoppable hunger for commercial growth on the other.

For now and thanks to his own personal reasons for wanting to chase AGI, Hassabis was keeping these battling ideologies at arm's length. He was in England, living thousands of miles away from the Silicon Valley bubble, and he had surrounded himself with a team of spectacularly clever AI scientists and engineers, a team that was about to grow even larger. Hassabis resolved that he would crack the conundrum of AGI in the next five years, most likely earning a Nobel Prize along the way, according to people who worked with him. It didn't matter that he was being folded into a corporate Goliath. Once he had built AGI, the concept of economics would become antiquated, and DeepMind and Google wouldn't have to worry about making money. AI would solve that problem away.

When the deal was finally inked and the ethics board added to the acquisition agreement, Google was buying DeepMind for $650 million. It was considerably less than what the founders would have gotten from Zuckerberg but a huge amount of money for a British technology company, and it came with that all-important agreement to keep control of AGI out of the hands of a large corporation.

Google's influx of cash also meant Hassabis could headhunt even more talented researchers. While some staff didn't like selling out to Google, many were elated at their huge pay rises and even more lucrative stock options in Google, which made them much less likely to jump ship to other technology firms. Now instead of worrying about Facebook or Amazon poaching his staff, Hassabis could poach *their* staff and lure some of the greatest AI minds from academia with eye-popping salaries. As he got the company on track to make even more advanced technology, Hassabis maintained DeepMind's secretive culture, to the point that the company's main website remained just a blank page with a circular logo in the middle. So mysterious was the AI lab that when people applied for jobs at DeepMind's headquarters in London, staffers wouldn't even put the company's address in an email. A representative would meet the candidates at the nearby Kings Cross train station and lead them to the office on foot.

In job interviews, the founders were persuasive, Suleyman especially, according to one former executive: “He was hugely charismatic, communicating that this is a once-in-a-lifetime opportunity to be part of something that was going to change the world.”

Academics and civil servants who’d spent a decade or more in their careers, and who could have easily gone into other high-paying roles in the private sector, would walk out of a twenty-minute conversation with Suleyman convinced that they should help build AGI. “He explained that the revolution would be built on better math,” the former exec adds. Hassabis and Suleyman would say they were hiring “the best mathematicians and physicists in the world.” And now thanks to being part of Google, they had access to the world’s best supercomputers and the most data for training AI models too.

About 50 percent of DeepMind’s recruits were now coming from academia, and they could hardly believe their luck. They’d gone from being squashed up against filing cabinets and begging for grant money to a place with gleaming offices in the middle of cosmopolitan restaurants and gardens, boasting ultrafast computers and virtually unlimited resources. And the best part was that DeepMind made sure you didn’t feel like you were working for an advertising giant. You were conducting research at a prestigious scientific organization that published papers in peer-reviewed journals like *Science* and *Nature* and solving the world’s biggest problems. It was the best of both worlds, if such a thing were possible.

In the long run, it wasn’t. But the six-figure salaries and unbelievable benefits made DeepMind staffers forget how odd it was to be getting paid so spectacularly well by Google to simply make the world a better place. Occasionally these moments of incongruity would poke through, such as when old colleagues from the grubby world of academia or the civil service would ask to come and visit.

“I used to be embarrassed,” says one former staffer, who had joined DeepMind from an academic job. When his former coworkers asked if they could see his new office, the staffer talked them out of it. He suggested visiting a nearby restaurant instead. Even that would be a little more modest than the DeepMind canteen, which served the kind of five-star buffet you’d find in a Dubai hotel. “It felt so separate from the real world,” they add. “It was actually ridiculous.”

Researchers were treated like rock stars and waited on hand and foot. One of them once emailed DeepMind’s staff-support service, which was normally used for expenses or getting visas, to say that it’d be a more efficient use of time if all the strawberries had their leafy caps removed. Two days later, the buffet had bowls of gleaming, hulled strawberries with not a speck of green.

Staff were endlessly reminded about the vision of building AGI, with Hassabis often telling them that at their current rate of research and breakthroughs, their end goal was just five years away. Hassabis was a master at painting an inspiring vision of where the company was heading, according to people who worked at DeepMind. At team off-site gatherings, he and Suleyman would give presentations about strategy that felt more like pep rallies than an explanation about specific future steps. The founders often didn't go into granular detail about tactics.

"It was all very vision-led, and 'Let's get on board with the mission,'" says a former staffer. "Demis and Mustafa were extraordinary, amazing storytellers. They balanced each other incredibly well." Hassabis was the serious brain who read scientific papers late into the night, who talked through methodologies for hours with his top researchers, and who also tended not to consort with lower-ranked staff who didn't have PhDs. It was Hassabis who fashioned a deeply hierarchical culture at DeepMind that was largely based on academic repute. Suleyman was the charismatic visionary when it came to rendering a vision of the future that everyone was working toward. One former staffer says he was like DeepMind's pied piper. Legg, the most academic of the trio, faded into the background somewhat. "Shane was quieter," the staffer says.

Hassabis believed so fervently in the transformative effects of AGI that he told DeepMind's staff they wouldn't have to worry about making money in about five years, because AGI would make the economy obsolete, former employees say. That eventually became mainstream thinking among the senior managers. "They'd drunk their own Kool-[A]id," says one former executive. They thought: "We are creating the technology that is the most important that humans have ever seen."

Behind the scenes, Hassabis and Suleyman were putting together the ethics and safety board that Google had agreed to as a condition for buying DeepMind because they knew they needed a fail-safe, with Suleyman its main advocate. Google had a fiduciary duty to its shareholders to grow its profits every year, which it continued to do very successfully. While that gave DeepMind the talent and computing resources it needed to build AGI, the situation was a double-edged sword. When they *did* create AGI, Google would almost certainly want to monetize and control it. They weren't sure how, exactly, but the board would at least make sure their human-level AI wouldn't be misused.

About a year after the acquisition, DeepMind convened its first meeting for the ethics and safety board at a conference room inside SpaceX's headquarters in California. Hassabis, Suleyman, and Legg were on the board, and so were Elon Musk and Reid Hoffman, the billionaire cofounder of LinkedIn turned venture capital investor. The other men at that first meeting included Larry Page, Google executive Sundar Pichai, Google's legal chief Kent Walker, Hassabis's postdoc advisor Peter Dayan, and Oxford University philosopher Toby Ord, according to people with knowledge of the meeting.

The meeting went well, but then the founders got some surprising news from Google. The company didn't want its new ethics board to go forward after all. Suleyman was angry, since he'd pushed for the board's establishment. Part of Google's explanation at the time was that some of the board's key members had conflicts of interest—Musk was potentially backing other AI efforts outside of DeepMind, for instance—and establishing a board just wasn't legally feasible. To some of the board's short-lived members, that sounded like baloney. They suspected that in reality, Google just didn't like the idea of being at the mercy of a group of people who could take away its control of lucrative AI technology.

Angered by what felt like a betrayal of Google's agreement, Hassabis and Suleyman complained to company leadership about losing the board. The executives needed to keep DeepMind's founders happy and pushing the boundaries of AI research, and they found a way to dangle a bigger prize in front of them. A senior Google executive reached out to Hassabis and his cofounders and told them there might be a better structure they could use to protect their AGI tech. The DeepMind founders didn't know this at the time, but Google was preparing to turn itself into a conglomerate called "Alphabet," which would allow its various business divisions to operate with more independence. The executive told the founders that these new divisions would be called "autonomous units." It would be like becoming an independent company again. They would get their own budgets, balance sheets, boards, and even outside investors. The idea sounded promising.

⋆ ⋆ ⋆

Out of view, Google's real goal was to boost its share price, which had been stagnating. For years, Wall Street analysts had been struggling to evaluate Google's bundle of other businesses outside of YouTube, Android, and its lucrative search engine. It had all these other businesses, too, like a smart thermostat company called Nest, a biotech research firm called Calico, a venture capital unit, and the "moonshot" X lab. Most of these divisions didn't make any money, but if they were turned into separate firms housed under a parent company, that could loosen up the company's balance sheet and help enhance the value of the business that Google cared about most: advertising. Google's advertising business made up more than 90 percent of its annual revenue. Despite its reputation as an innovative technology company filled with the brightest engineers, Google's leadership still primarily cared about the age-old business of getting people to buy stuff they didn't necessarily need.

With so much energy devoted to building AGI, Hassabis, Legg, and Suleyman barely stopped to think about Google's true motivations, or that it probably had no intention of ever giving them autonomy when their artificial intelligence research could be so useful to growing its business. Instead, the idea of becoming more independent was music to their ears. It meant Google wouldn't control their future AI, and *they* could be its careful stewards. "We wanted to have sufficient independence to be able to navigate what might be coming if very powerful AGI was coming along," Legg remembers. "We wanted to make sure we had enough control over how things developed."

The founders spent the next year and a half talking to Page and other executives about how their existence under this new corporate umbrella might look and what "autonomous unit" actually meant. But then, when Google announced that it was being restructured under the name Alphabet, it wouldn’t confirm or announce any plans to give DeepMind more legal autonomy. As several other Google bets, like Verily Life Sciences, got spun out as separate companies, there was no progress on doing the same for DeepMind. It was almost as if Google had forgotten about its commitments yet again.

Hassabis didn’t have much time to dwell on the way Google seemed to be fobbing him off. There was a more troubling matter coming up on the horizon. Over in San Francisco, some start-up founders were setting up another research lab that had the same goal as DeepMind’s. They were touting a big, new idea to build artificial general intelligence safely and for the benefit of humanity. The implication stung a little bit, suggesting that the world’s other big attempt to build AGI—his own—wasn’t helping humanity. It was helping Google. What made things worse was that this new organization had been spun up by his old investor, Elon Musk. It was called OpenAI.

CHAPTER 6

The Mission

It was 2015, and for five years, Demis Hassabis had been growing his team and hitting research milestones on a slow but steady path to AGI, operating in an open field with virtually no one else trying to do the same thing. DeepMind's objective was so radical that it could effectively operate like a monopoly. No other established company in the world was trying to build AI that could surpass human intelligence, which meant Hassabis could conduct research at his own pace. That also made it easier for DeepMind's founders and staff to see themselves as more of a mission-oriented research lab than a company. They could mentally reconcile themselves with being owned by Google but still "solving intelligence" to fix humanity's biggest problems because they weren't running on the same never-ending hamster wheel of competition that other companies were. Their quest was unique. Now the possibility of a rival in Silicon Valley was going to change all that. The quest to build AGI was about to turn into a race.

The more Hassabis learned about OpenAI, the more his anger rose. He had been the first person in the world to make a serious run at building artificial general intelligence, and given what a fringe idea it had been five years earlier, he'd put his neck on the line with the scientific community by doing so. To make matters worse, this new contender might even be exploiting his ideas. OpenAI had seven people listed as cofounders on its website. When Hassabis took a closer look at the names, he realized that five of them had worked as consultants and interns at DeepMind for several months. That's when he became livid, according to people who worked with him. Hassabis had been an open book with DeepMind staff about the different strategies they needed to chase to reach AGI, such as building autonomous agents or teaching AI models to play games like Chess and Go. Now five scientists who'd heard all those details were starting a competing organization.

Technically speaking, Hassabis might not have had to worry that much. There were many other researchers outside of DeepMind who were doing similar work with autonomous agents, virtual environments, and games. One of those five former visitors was a renowned AI scientist named Ilya Sutskever, who specialized in deep learning, not DeepMind's signature technique, reinforcement learning. Sutskever was OpenAI's chief scientist and, like his cofounders, a deep believer in the possibilities of AGI.

But Hassabis still bristled at Sam Altman's audacity in hiring people who knew of DeepMind's secrets, and anxieties crept up on him late at night. Hassabis would typically come home from work to have dinner with his family before embarking on the second part of his working day, starting in the evening and going on till 3:00 or 4:00 a.m., reading research papers and sending emails. Hassabis worried aloud in some of those emails or late meetings that Altman was copying DeepMind's strategy and trying to steal its researchers, according to people who heard those comments.

Hassabis questioned OpenAI's promises to release its technology to the public. That approach to being "open" seemed reckless. "I thought it was a bit naive that open-sourcing was a panacea," he says today. "As you get more and more powerful dual-purpose technologies, what about bad actors accessing that technology, for bad ends?… You have very limited control over what somebody might do." DeepMind published some of its research in well-known journals, but it kept the full details of its code and AI technology under tight control. It didn't release the AI models it had created to master the game *Breakout*, for instance.

Deepening the humiliation, DeepMind leaders caught wind that Musk was trash-talking Hassabis to his contacts in Silicon Valley, according to people who worked at DeepMind and OpenAI. When the billionaire was talking to all the new staff at OpenAI, for instance, he warned them about DeepMind's work in England and suggested Hassabis was a shady character. He cast suspicion over the way Hassabis had designed *Evil Genius*, a game where you played a villain trying to build a doomsday device and dominate the world. Whoever created games like that was probably a little maniacal themselves. OpenAI's staff ran with the joke and created memes based on screenshots of *Evil Genius*, which they would send to one another on Slack, the chatting service. At one point, Musk referred to Hassabis as the "Hitler of AI," according to a former OpenAI staffer who heard the comment directly.

Whatever his reason for turning on DeepMind, Musk was stoking what would become an intense rivalry between the two organizations. He'd also been picking up a more paranoid, pessimistic view of AI that tracked with his tendency to take things to their extreme. He could have, for instance, simply fought oil companies to tackle climate change, but decided to make humans an interplanetary species instead. He could have bought a stake in Twitter when he resolved it was too woke, but he bought the whole company. Maybe it was Musk's habit of taking drastic action, his tendency to exaggerate, or his belief in his role as humanity's savior, but within a couple of years of investing in DeepMind, the tycoon was tunneling deep into the dogma of AI doom.

He'd been having late-night conversations about the issue with his wife, worrying about how Google's quiet cofounder Larry Page was on course to make far more advanced AI systems after buying his former investment DeepMind, according to reporting by the *New York Times.* Musk and Page were close friends. They went to the same exclusive dinners and conferences and had similar fantastical dreams about the future. If Musk was in San Francisco and hadn't arranged for a place to stay, he'd call

Page and ask if he could crash on his sofa, according to a biography of Musk by Bloomberg reporter Ashlee Vance. They'd play video games and bat ideas around for futuristic airplanes or other technology. Musk thought that Page, who was becoming more and more reclusive, was almost too nice. That started to worry him. Google's cocreator might produce something evil by accident, Musk said in his biography, like a "fleet of artificial- intelligence-enhanced robots capable of destroying mankind." Musk sounded like he was joking, but he was serious.

Months after Page bought DeepMind for $650 million, Musk posted and then quickly deleted a message on a web forum about AI. Nobody realized how quickly AI was developing, he said. "Unless you have direct exposure to groups like DeepMind, you have no idea how fast." He said he was skeptical that certain "leading AI companies" could prevent digital superintelligences from escaping into the internet and causing havoc.

As Musk went down the rabbit hole of AI doom, he started investing more of his money and time in the issue. He gave $10 million to the Future of Life Institute, a nonprofit organization that campaigned for more research into stopping human annihilation through AI. Then when the group set up a conference in Puerto Rico, he attended, along with Larry Page, Hassabis, and anyone else who was serious about building AGI.

After one dinner party at the conference, Musk and Page got into an argument. As it got heated, more conferencegoers started to surround them to listen: Page said that Musk was becoming way too paranoid about AI. He had to remember humanity was evolving toward a digital utopia, where our minds would become digital and organic. If he kept making such a fuss about AI, he'd slow down all the next steps there.

"But how can you be so sure that a superintelligence won't wipe out humanity?" Musk asked.

"You're being speciesist," Page shot back, according to the *New York Times* account, apparently defensive of the future's posthumans. By focusing so much on catastrophe, Musk was dismissing the needs of all those future beings destined to be made from silicon.

On the one hand, as he kept tabs on DeepMind and immersed himself in a wealthy community of futuristic prognosticators, Musk was becoming radicalized. But on the other hand, he was also experiencing FOMO, the debilitating "fear of missing out" that fuels some of the biggest decisions in Silicon Valley about where to put money. As AI hit new milestones, such as the 2012 ImageNet competition victory, the large tech firms were sitting up and paying more attention to the field. Not only had Google bought DeepMind, but Mark Zuckerberg had set up a new division called Facebook AI Research, or FAIR, and hired one of the world's leading specialists in deep learning, Yann LeCun, to run it. It was likely this desire to be part of this new gold rush for research that led Musk to do something so counterintuitive to his fears: create more AI.

Later, Musk would say on Twitter that he had started OpenAI because he wanted to create a "counterweight to Google" and because he wanted AI to be developed more safely. But there was no doubt that AI was critical to the financial success of his companies, whether it was the self-driving capabilities of Tesla cars, the systems steering SpaceX's unmanned rockets, or the models underpinning his upcoming brain-computer interface company Neuralink.

For all of Musk's apocalyptic views and moral convictions that he should reach AGI before Demis Hassabis, building AI that was as capable as Google's would also boost his businesses. It was a profitable endeavor. Only that could explain why he agreed to work on that with one of the best-connected entrepreneurs in Silicon Valley: Sam Altman, the guy who turned "millions" into "billions" on slide decks, the guy who'd stuffed Y Combinator with futuristic start-ups, and the guy whose ambitions for AI were as big and far-reaching as Larry Page's.

Altman sent Musk an email on May 25, 2015, saying that "someone other than Google" should build AGI first. He suggested an AI project structured "so that the tech belongs to the world." "Probably worth a conversation," Musk replied. A month later, Altman emailed again, proposing a lab that would build "the first general AI.… Safety should be a first-class requirement." The AI would be owned by a nonprofit and used "for the good of the world." Musk replied, "Agree on all."

For Altman, building an all-purpose AI system was like taking all the technology start-ups he'd ever mentored in Y Combinator and putting them into one big Swiss Army knife. This powerful machine intelligence could be infinitely capable. Who knew if we'd even need businesses or start-ups anymore when a new superintelligence could generate enough wealth to keep everyone on Earth economically thriving? While Hassabis had believed that AGI would unlock the mysteries of science and the divine, Altman would say he saw it as the route to financial abundance for the world. He and Musk talked about starting a research lab that could do just that, and act as a counterbalance to DeepMind and Google.

Musk and Altman decided on another way their new organization would be different to Big Tech firms. In its effort to build AI that was good for humanity, it would collaborate with other institutes and make its research open to the public. Hence the name: OpenAI.

Altman got to work on establishing an initial founding team. That summer in 2015, he invited about a dozen top AI researchers to dinner at a private room at the Rosewood, a luxury hotel that was walking distance from some of Silicon Valley's wealthiest venture capital firms. The invitees included Ilya Sutskever, the scientist who'd spent several months at DeepMind, and Greg Brockman, a Harvard math graduate from North Dakota who was a whiz at building businesses and had been chief technology officer at Stripe.

During the dinner, Altman explained that the goal of this new research organization would be to build AGI and then distribute its benefits to the world. The group spent a large chunk of the meal asking if that was even possible—not the part about distributing

AI’s riches to humanity, but the one about starting such a lab when Big Tech firms had poached most of the world’s top AI talent. Wasn’t it too late to try to hire the field’s best researchers?

“We [also] knew our resources would pale in comparison to the [Big Tech] companies,” Brockman later recalled on the Lex Fridman Podcast. But if they did start such an organization, how should it be structured to make sure its AI benefited humanity? “It was clear that such an organization needed to be a [nonprofit], without any competing incentives to dilute its mission.”

Halfway through the drive home in Altman’s car, Brockman declared that despite how unrealistic this was, he was on board. This was Silicon Valley after all, where even the craziest ideas found a way to thrive.

Being a workaholic himself, Altman was impressed at how Brockman immediately started planning all the necessary logistics for setting up OpenAI. This was a guy whose average response time to emails was five minutes, which meant he could be as freakishly dedicated to the cause as Altman was. “He was fully in,” Altman later said. When it came to building OpenAI, Brockman would become the organizer for everything.

Brockman then took charge of poaching an initial group of talented scientists from companies like Google and Facebook, reaching out to Yoshua Bengio, a professor from the University of Montreal who’d been called one of the “godfathers” of the deep-learning movement. Brockman didn’t want to hire Bengio. He wanted the professor to tell him who were the most promising scientists in AI that he could think of. Bengio typed out a list of names and sent it back to Brockman.

Hiring these people wouldn’t be so easy. Some of them were earning seven-figure salaries with companies like Google and Facebook, and Altman and Brockman couldn’t offer anywhere near those amounts. What they did have was a compelling mission to change the world and two prestigious names running the show. Elon Musk was now a globally revered tycoon, and running Y Combinator had elevated Altman’s status in the Valley to someone everybody wanted an introduction to. For AI researchers, even a short stint at this new nonprofit group offered prestigious connections and a potential career boost that could be worth the pay cut.

Several of the leading scientists on Brockman’s list decided to meet him about the job. Alongside the big names and vision, they liked the “open” part of this new organization. They would get a chance to finally *publish* their research instead of working in secret on some corporate product, and some also liked the idea of counteracting the profit motive that was driving Google and DeepMind’s efforts to build AGI, according to former OpenAI staff.

To seal the deal, Brockman took a handful of scientists to a winery. Sutskever would be the biggest catch of all if he agreed to join. The group talked more about building an AI lab that would be completely free from corporate pressures and that would “open-source” its research, effectively giving it away for free, and how that would stop Big

Tech firms like Google and Facebook from having a stranglehold on AI as it became more powerful. Nearly all of the scientists agreed to come on board, including Sutskever, the talented scientist who rarely seemed to smile. Having grown up in Russia and Israel and having worked with prestigious deep-learning pioneer Geoffrey Hinton, he would now be ditching Google Brain for OpenAI.

With about a dozen people on board, the team headed to Montreal, Canada, for an annual AI conference called NIPS (now called NeurIPS) in December 2015, to announce the new research lab. Snow was building up outside the venue as members of the team talked to other conference participants about their new lab. The real announcement happened online. A website, OpenAI.com popped up with a blog post written by Brockman and Sutskever introducing the project. "Our goal is to advance digital intelligence in the way that is most likely to benefit humanity as a whole, unconstrained by a need to generate financial return," they wrote.

Musk and Altman would chair the organization, and it had an eye-popping $1 billion in funding commitments from Musk, Thiel, Altman, Hoffman, and Jessica Livingston, along with cloud computing credits from Amazon. Musk was planning to fund OpenAI with Tesla stock, according to a person with knowledge of the plans, just as he'd offered to do with DeepMind several years earlier.

The hundreds of academics at NeurIPS were astonished as they took in the news. Many thought that building AGI was a pipe dream, but some were also envious. For the past decade, Big Tech companies had been siphoning universities of their top computer science talent, and it was getting to the point that the very best minds in artificial intelligence were now working for corporate interests. Effectively, there was now an assembly line in artificial intelligence, which started at elite universities and ended at Google, Facebook, and Amazon. It had been a problem for years.

"There's no way that anybody would say 'no' to two or three times their salary," says Maja Pantic, a computer science professor at Imperial College London who joined Samsung Electronics in 2018 as research director of its AI center and then hopped over to Meta. "That is what happened to me. This is what happened to all of my colleagues." And to the luminaries too. Hinton now worked for Google; Fei-Fei Li left Stanford for Google; LeCun, for Facebook. Ng left Stanford for Google and then China's Baidu. Even the top universities like Stanford, Oxford, and the Massachusetts Institute of Technology could barely hold on to their star academics, leaving a vacuum where the next generation of educators was meant to be. AI research became more secretive and more geared toward making money. That's why Musk and Altman's push for their research to be open to the public was so refreshing to researchers. Someone was finally addressing the concentration of AI knowledge happening at big companies.

The university brain drain was happening for two reasons. The first and most obvious was pay. At the University of Toronto, where Geoffrey "Godfather of AI" Hinton had once taught, computer science professors could expect to earn about $100,000 a year. The university's highest-earning scholars were bringing in about

$550,000. That was the top of the range. Hinton's star student Sutskever didn't even try going into academia. After a stint at Hinton's start-up, he'd gone straight to Google Brain. When OpenAI offered Sutskever $2 million a year to join, Google Brain offered three times that amount, according to *Genius Makers*.

A second reason was the data and computing power needed to run experiments in AI research. Universities typically have a limited number of GPUs, or graphics processing units, which are the powerful semiconductors made by Nvidia that run most of the servers training AI models today. When Pantic was working in academia, she managed to purchase sixteen GPUs for her entire group of thirty researchers. With so few chips, it would take them months to train an AI model. "This was ridiculous," she says. Not long after she joined Samsung, she got access to two thousand GPUs. All that extra processing power meant that training an algorithm could take days, and their research could speed ahead.

For those scientists who stayed in academia, it was also becoming harder to escape Big Tech's sway. One 2022 study found that over the previous decade, the number of academic papers that had ties to Big Tech firms had more than tripled to 66 percent. Their growing presence "closely resembles strategies used by Big Tobacco," said the authors of the study, which was conducted by researchers across several universities, including Stanford and University College Dublin. That in turn influenced the way universities measured success with their AI research. Instead of aiming for values like the well-being of people, justice, and inclusion, academics were more inclined to aim for better performance, according to Abeba Birhane, now a senior fellow at the Mozilla Foundation, who led the study.

Well-being and inclusion weren't just wishy-washy concepts, says Birhane. They were perfectly measurable. "They may be abstract, but so [are] efficiency and performance," she adds. "People have found ways to measure fairness, privacy and more." What was making the matter worse was that as researchers everywhere, from universities to tech companies, focused so much on making their AI models bigger and more capable, they were also raising the risks that those models could sometimes produce outputs that were racist or sexist, says Birhane, who points to another 2023 study that she coauthored. "What we found was that no, as your dataset scales, hateful content also increases."

Yet scale was critical to the growing power that large tech companies were amassing in AI. Google and Meta, the company formerly known as Facebook, had trillions of data points that could be used to train models, and they ran server farms that spanned hundreds of thousands of square feet. A single data center that Google currently runs in Dalles, Oregon, for instance, is bigger than six football fields. Most universities can offer only a tiny fraction of that.

When it came to making AI smarter, more was better. As he kicked off his research at OpenAI, Sutskever and his team focused on making AI models that were as capable as possible, not necessarily as equitable, fair, or private. In very simple terms, there was a formula for doing that. If you trained an AI model with more and more data, and you also raised the number of parameters the model had, and you *also* boosted the computing power used for training, the AI model would become more proficient. It was the same extraordinary correlation that Professor Andrew Ng had noticed when he was doing his experiments at Stanford. It didn't matter what your model was designed to do. So long as you turned up all the dials, it would be more accurate at translating language, or it would sound more human when it generated text.

"If you have a very large dataset and a very large neural network, success is guaranteed," Sutskever said at one AI conference. The last three words of that statement became his catchphrase among AI scientists, all the more so after OpenAI's big launch, as the field took on a new air of excitement about this new nonprofit led by a brilliant scientist and several of Silicon Valley's biggest power brokers.

It wasn't long before problems started to arise. OpenAI didn't immediately get the $1 billion in funding commitments it had announced in December, from Musk, Thiel, and others. In fact, over the next few years, the nonprofit managed to collect only a little over $130 million in actual donations, according to an investigation by the tech news site TechCrunch, which pored over OpenAI's federal tax filings.

OpenAI lacked money and it was hazy in its direction. Its founding team of thirty researchers started working out of Brockman's apartment in San Francisco's Mission District, at his kitchen table or slouched on sofas with their laptops perched on their knees. A few months after the launch, they got a visit from another respected Google Brain researcher named Dario Amodei. He started asking some probing questions. What was all this about building a friendly AI and releasing its source code into the world? Altman countered that they weren't planning to release all the source code, according to his *New Yorker* profile.

"But what is the goal?" Amodei asked.

"It's a little vague," Brockman admitted. Their goal had been to make sure AGI went well.

Amodei was part of the growing cohort of scientists who had similar fears of doom to Musk and Eliezer Yudkowsky. He'd been working at Google when, less than a year earlier, the company came under fire after the vision recognition system in its Photos app was spotted classifying people of color as gorillas. Google said it was "appalled" and removed the gorilla label from Photos completely. "Having systems that fail unpredictably is not a good thing," he told one podcast, referring to the incident.

But Amodei's concerns weren't limited to racist and offensive decisions by algorithms. He was also worried about how reinforcement learning, the AI technique being mastered by DeepMind, was being used to control physical systems like robots, self-driving cars, and Google's data centers. "Once you're actually interfacing with the world directly and controlling direct physical things, I think the potential for things to go wrong … starts to increase," he said in a 2016 interview with Jaan Tallinn's Future of Life Institute.

Amodei's research into AI harms led him to more and more catastrophic possibilities, and by 2023, he would be warning the media that there was a 25 percent chance runaway AI posed an extinction risk to humans. Google Brain wasn't where he could tackle such risks. After that probing conversation at OpenAI's office, he joined a few months later.

To build AGI, OpenAI's founding team needed to attract more money and talent, so they tried focusing on projects that could generate positive stories in the press. Their early researchers created a computer that could beat the top human champions at *Dota*, a strategic 3D video game, and they also built a five-fingered robotic hand, powered by a neural network, that could solve a Rubik's Cube. These projects were aimed at keeping Elon Musk happy by trying to one-up the work happening across the Atlantic in the secretive offices of DeepMind.

Musk didn't keep his mistrust of DeepMind a secret. In 2017, OpenAI's staff went to an off-site meeting at SpaceX's headquarters. Musk, who'd been visiting the OpenAI office every week at first and then every few weeks, gave them a tour of the facilities and then did a question-and-answer session with about forty of his new artificial intelligence researchers. At one point, Musk started talking about why he'd funded OpenAI, and the reason was Demis Hassabis.

"I was one of the investors in DeepMind, and I was very concerned that Larry [Page] thinks Demis works for him. Actually, Demis just works for himself," Musk said, according to a person who was there. "And I don't trust Demis."

The researchers were astonished. To many of them, it sounded like Musk had a personal issue with Hassabis more than any particular worry about where AI was headed. When he was asked about his antagonism for Hassabis, he mentioned the computer games that the British entrepreneur had designed in the past that focused on world domination.

At the same session, Musk recounted a conversation he'd had with another investor in DeepMind, who'd said that during an earlier meeting with Hassabis, "I felt like it was at that point in a movie where somebody should get up and shoot the guy." In other words, somebody needed to stop Hassabis from building an all-powerful AGI.

But as much as Musk didn't seem to like Hassabis, he would remind OpenAI staff that DeepMind was ahead, and he held up the British company's research work as a benchmark they needed to target. As the months passed, Musk grew more and more concerned that OpenAI's technology simply wasn't as powerful as DeepMind's, according to former OpenAI staff.

To keep their biggest benefactor on board, Altman and Brockman steered some of their researchers toward emulating the work that DeepMind was doing. Researchers on the *Dota* project, for instance, couldn't understand why they were working on a game simulation if their ultimate goal was to build an AGI that would make people's lives better. The reason was they needed Musk's money. "If we don't work on this, OpenAI might not exist in a few years, or even next year," Brockman told the researchers.

Although OpenAI eventually gained worldwide acclaim for its work on chatbots and large language models, its first few years were spent toiling on multiagent simulations and reinforcement learning, fields that DeepMind already dominated. But the more they chased DeepMind in those fields, the more Altman and his leadership team realized that these approaches to AI didn't promise much real-world impact. That's when OpenAI started to evolve into a very different kind of organization to DeepMind. While DeepMind had a hierarchical, academic culture that prized its PhD staff, OpenAI's culture was more engineering-led. Many of its top researchers were programmers, hackers, and former start-up founders at Y Combinator. They tended to be more interested in building things and making money than in making a discovery and achieving prestige within the scientific community.

Musk, meanwhile, was getting antsy. He complained to Altman that he had recruited an impressive roster of scientists but didn't have any demos that blew DeepMind out of the water. As the nonprofit approached its third year, Musk told Altman that it was falling too far behind Google and DeepMind. He then offered a quick solution: he would take control of OpenAI and merge it with Tesla. OpenAI would never catch up to DeepMind without a major change, Musk said in a December 2018 email to Altman and his team, which was published by OpenAI and corroborated by someone who saw an unredacted version. "Unfortunately, humanity's future is in the hands of Demis," he added. In other words, Hassabis the villain would get his way if Musk didn't take charge. But Altman and his cofounders wanted to stay in control. They rejected Musk's proposal.

In February 2018, OpenAI briefly mentioned in a public announcement about new donors that Musk was leaving, but it framed the reason as benign. Musk was leaving for ethical reasons. He had too big a conflict of interest in the field of AI. "Elon Musk will depart the OpenAI board but will continue to donate and advise the organization," the nonprofit said on its blog. "As Tesla continues to become more focused on AI, this will eliminate a potential future conflict for Elon."

Many staff at OpenAI knew that was hogwash. They suspected that as much as Musk said he cared about creating safer AI, he also wanted to be the person who built the most capable AI. He was already the wealthiest man on Earth and gaining unprecedented sway over American infrastructure: NASA was putting astronauts into space with SpaceX; Tesla was leading the charge on electric vehicle standards; and Musk's satellite internet company, Starlink, was on course to try to shape the outcome of the Ukraine war.

It was clear that Musk was also chronically unreliable. He had promised to donate $1 billion to OpenAI over several years, but instead had put in somewhere between $50 and $100 million— a rounding error for the world's richest worrier about AI. Putting in that money would have been relatively easy, especially if he was going to fund OpenAI with Tesla stock. Between 2015 and 2023, shares in Tesla jumped more than 18,000 percent, meaning OpenAI could have hit that $1 billion funding target without too much trouble. For all the concerns Musk had about humanity's future, he seemed far more preoccupied with staying ahead of the competition.

As Musk left OpenAI, he took its main source of funding with him. This was a disaster for Altman. He had staked his entire reputation on this project. Some of the world's top AI scientists were on a pay cut just to work with him, and his grand promises to help humanity were starting to look silly. The simple truth in this new era of AI development was that you needed more of everything to be successful, from money to pay your researchers to data to train your models and powerful computers to run them. Without Musk, the chance of ticking those boxes was quickly diminishing.

Altman was approaching a critical juncture. Working out of OpenAI's office in San Francisco, he thought about how he could keep the nonprofit going on severely limited resources and build AI models that were likely to be subpar to the rest of the field. The other option was to call it a day and close the project down. Raising money for a nonprofit was much harder than fundraising for a start-up. Altman was struggling to convince wealthy people to donate to the cause of AGI out of the goodness of their hearts, with no chance of seeing a direct financial return. He needed tens of millions of dollars and Musk had been his last big benefactor.

There was another option. Maybe OpenAI could give its backers some kind of direct economic benefit, in addition to the honor of sparking an AI utopia for humanity. That would be a win-win. The backers wouldn't be making a "donation" so much as an "investment," which was the language Altman was more comfortable speaking anyway. But he saw only a few potential backers he could realistically approach to get both the money and computing power OpenAI needed to build AGI. They were tech giants like Google, Amazon, Facebook, and Microsoft. No one else had the billions of dollars on tap or powerful computers housed in buildings that spanned football fields.

For the last few years, both OpenAI and DeepMind had been trying to put roadblocks in place to stop whatever ultrapowerful AI system they made from being misused. DeepMind was trying to change its governance structure so that a profit-motivated monopoly in the form of Google wouldn't have free rein to monetize AGI. Instead, a council of expert advisors would keep things in check. Altman and Musk had established OpenAI as a nonprofit and promised to share its research and even its patents with other organizations if it looked like they were getting closer to the threshold of superintelligent machines. That way it would prioritize humanity.

Now as Altman fought to stay alive, he was going to knock down some of those guardrails. The cautious approach he'd started with was going to morph into something more reckless, and doing so would transform the AI field that he and DeepMind had been working in from a slow and largely academic pursuit into something more like the Wild West. Altman would use his ability to spin a compelling narrative to justify the departure he was about to take from OpenAI's founding principles. He was a tech founder, and tech founders had to pivot sometimes. That was how it worked in Silicon

Valley. He would only need to *tweak* some of OpenAI’s founding principles—just a little bit.

ACT 2

THE LEVIATHANS

CHAPTER 7

Playing Games

A short walk from London's King Cross Station, where tourists flocked to see the magical train platform that Harry Potter took to Hogwarts, a different kind of magic was being created inside a gleaming collection of high-rise buildings jutting into the gray sky, their facades a blend of glass and metal cladding. Between them was a pretty promenade bustling with pedestrians. Some of them were the engineers and artificial intelligence scientists of DeepMind, fishing badges out of their pockets as they walked through the glass doors of an office building that officially belonged to Google but that had two floors devoted to their secretive AI lab.

For all the perks that DeepMind got from being part of Google, including the nap pods, massage rooms, and indoor gym, its founders were still trying to extricate themselves from the grip of their parent company, Alphabet. It had been more than two years since the acquisition, and the tech giant's executives were dangling a new prospect in front of Demis Hassabis, Mustafa Suleyman, and Shane Legg. Instead of being an "autonomous unit," DeepMind could become an "Alphabet company" with its own profit-and-loss statements.

Being in England and far removed from the relentless growth ethos that drove Silicon Valley, the founders took Google's suggestion in good faith. Suleyman wanted to show that DeepMind could stand on its own two feet as a business, so he dove into proving out the value of its AI systems in the real world. He put renewed focus on a division he'd started called Applied, whose researchers used reinforcement learning techniques to tackle problems in healthcare, energy, and robotics to potentially turn into businesses. Another team of about twenty researchers, who called themselves DeepMind for Google, worked on projects that directly helped Google's business, making YouTube's recommendations more efficient, for instance, or improving Google's ad targeting algorithms. Google agreed to give DeepMind 50 percent of the proceeds of the value that it added to those features, according to someone with knowledge of those agreements. About two-thirds of the projects ended up being useful to Google, another former staffer says.

That left hundreds of other researchers at DeepMind free to continue researching ways to build AGI. Once every few weeks, the founders would meet at a London pub to talk shop, and their discussions would arrive at familiar points of tension. Suleyman wanted to solve real-world problems but also worried they could inadvertently build a superintelligent system that would go awry. What if the AI got out of its box and manipulated people? he asked. In the office he warned other staff and managers that AGI's impact on the economy could lead to a sudden displacement of millions of jobs and plummeting incomes. What if that led to an uprising? "People will be walking up to Kings Cross with pitchforks if we don't think about equality," he'd say, according to a former staff member.

Hassabis's brain would grapple for solutions, but they sometimes sounded a little off the wall. For instance, he'd suggest that as their AI got more powerful and potentially dangerous, DeepMind could hire Terence Tao, a professor at University of California, Los Angeles, who was widely regarded as one of the world's greatest living mathematicians. A former child prodigy who went to college at the age of nine, Tao had become known as a Mr. Fix-It for frustrated researchers, according to *New Scientist* magazine.

Tao had said in interviews that AI was largely clever mathematics and that the world would probably never have true AI. He saw the technology in the same mechanistic and almost black-and-white way that Hassabis did. If AI got out of control, math could contain it. Hassabis wasn't alone in believing that. On Yudkowsky's LessWrong forum, members had engaged in a lengthy project to brainstorm how they might convince someone like Tao and other top mathematicians to work on AI alignment, the practice of making AI more "aligned" to human values to prevent it going rogue. They tossed out figures of between $5 million and $10 million for how much these math luminaries would need to be paid.

Hassabis imagined that as he got close to AGI, he'd stop pushing the performance of his AI models and then invite some of the world's greatest minds to come and analyze them down to the minutest detail so that they could help figure out the best calculations to contain them. "Maybe we should start putting out the call, an almost 'Avengers Assemble' of mathematicians and scientists," Hassabis still says today.

Suleyman disagreed with his cofounder's approach, believing it far too focused on numbers and theory. He believed AI needed to be managed by people, not just clever math, to make it safe. As he and Hassabis debated the best strategy for containing AI, they got another update from Google's leadership about the plan to become an "Alphabet company." That idea wasn't going to work after all, the executives told them. Spinning out wasn't straightforward because as AI had become increasingly valuable to Google's business, the larger company needed DeepMind even more.

The founders felt like they were experiencing déjà vu, as Google reversed course yet again. But the executives told them not to worry, because they could still find a compromise. They now suggested a third option: DeepMind could do a kind of partial spinout and have its own board of trustees guiding its creation of superintelligent AI, but Alphabet would retain some ownership of the AI company. To show they meant it, Alphabet put the commitment in writing. Its management signed a term sheet in which Google pledged to give DeepMind $15 billion in funding as a kind of endowment over ten years to run independently, according to a person with direct knowledge of that agreement. Hassabis told multiple people at DeepMind that the term sheet had been signed by Sundar Pichai, the Google executive who in a few years would become Alphabet's chief executive officer. That meant Google was serious about its commitment this time.

A term sheet is a document that outlines the terms and conditions of a potential business agreement. It usually acts as a starting point for further negotiations, and it's not legally binding. Still, an agreement in writing holds more weight than a spoken one, and the DeepMind founders believed Google's pledge to set them free was real this time. They decided to throw themselves into refashioning DeepMind as a different kind of business and one that—like OpenAI—would have a formal structure that made them more like a charity than a business.

Hassabis and Suleyman hired investment bankers to work out the financial mechanics of a spinout, and they also hired two London law firms to draw up the legal plans for restructuring themselves as an independent organization. They took advice from a top corporate litigator in the UK who had shepherded deals for larger companies like Shell, Vodafone, and mining giant BHP Billiton.

They also planned a new leadership structure: Hassabis, Legg, and Suleyman would sit on an operating board alongside Alphabet CEO Larry Page; his cofounder, Sergey Brin; Google's then–product chief, Sundar Pichai; and three independent commercial directors. Decisions would be made by majority vote. Crucially, there would also be a fully independent board of trustees made up of six directors who would oversee DeepMind's compliance with its social and ethical mission. The names of those directors, as well as their decisions, would be made transparent to the public. Since those six directors would be steering some of the most powerful and potentially dangerous technology in the world, they needed to be high-caliber, trustworthy people. So DeepMind reached for the stratosphere, asking former president Barack Obama to become one of those directors, along with a former US vice president and a former CIA director. Several of these people agreed to take part, according to someone who was close to that work.

After consulting with legal experts, DeepMind decided it would not go down the same route that Sam Altman initially had by becoming a nonprofit organization. Instead, its founders contrived a completely new legal structure they called a global interest company or GIC. The idea was that DeepMind would become an organization that was more like a division of the United Nations, a transparent and responsible steward of AI for humanity's sake. It would give Alphabet an exclusive license so that any AI breakthroughs DeepMind made that could support Google's search business would flow to the technology giant. But DeepMind would spend the majority of its money, talent, and research on advancing its social mission, working on drug discovery and better healthcare or tackling climate change. Internally, they referred to the project as GIC.

Yet even as they sought to carve themselves away from Google, DeepMind was simultaneously helping bolster Google's business. Around the time Google's Larry Page was promising to help DeepMind spin out, he was looking to China as a new opportunity for expansion. As Google's business in the United States and other Western markets matured, China presented a unique opportunity. It was the world's most populous country, with more than 650 million internet users, almost double the entire population of the United States. And only about half the people in China who could be online *were* online, meaning China offered a vast, untapped market. China's middle class was growing, consumer spending was on the rise, and its gross domestic product was around $11 trillion, making it the second-largest economy in the world. It was a potential gold mine for any internet company.

But Google couldn't just waltz into China. In fact, in 2010 it had exited the country after accusing Beijing of hacking its intellectual property and the Gmail accounts of Chinese human rights activists. The Chinese government had demanded that Google censor searches about Tiananmen Square and other controversial topics for the Chinese Communist Party. Then it blocked access to Facebook and Twitter, setting up what became known as the Great Firewall. Google's leadership were cocky and believed this was all just temporary, because China's citizens would soon enough be clamoring for the slick, powerful services offered by Silicon Valley's web giants.

"In a long enough time period, do I think that this kind of regime approach will end?" Google's Eric Schmidt, the company's chairman at the time, asked *Foreign Policy* magazine in 2012. "I think absolutely."

Schmidt was wrong. Instead of wasting away, China's own internet sector boomed. Companies like Meituan, Baidu, and Alibaba became juggernauts as Chinese engineers who'd worked and started companies in Silicon Valley flew back home to build their own tech leviathans. A large number of engineers who had worked in Microsoft Research Asia were taking leadership positions at Chinese internet giants such as Alibaba and Tencent. Five years after Google had left, it was watching the country's market become ever more lucrative but had no clear way back in. China's rules around censorship hadn't changed. But Google was eager to tap both the growing consumer market and some of the innovative engineering ideas that were blooming in China. "We need to understand what is happening there in order to inspire us," Google's head of search, Ben Gomes, told *The Intercept* at the time. "China will teach us things that we don't know."

At around this time, a major leadership change took place at the top of Google. In 2015, Page and Brin took a step back from the company they'd founded to chase an array of personal interests outside of Google, from philanthropy to flying cars to space exploration. They appointed Pichai as the new CEO. Pichai was the well-regarded product chief who the DeepMind founders were planning to put on their new operating board once they spun out. But unlike Page, he didn't have much time, or likely much inclination, to help one of Google's most prized acquisitions get away. He and Schmidt were busy looking for creative ways back into China. At one point that year, it looked like they might get Beijing's approval to bring their app store back to the country, but nothing happened.

Then came a public relations opportunity that would put DeepMind at center stage. DeepMind had been training its AI models with games, and its latest program, *AlphaGo*, could play the two-player abstract strategy board game of Go. Originating in China more than 2,500 years ago, Go looked deceptively simple. It is played on a nineteen-by-nineteen grid board with a few handfuls of black and white stones. The players each take turns placing a stone on an intersection of the grid. The goal: capture territory on the board by surrounding empty points with your stones, and get your opponent's stones too. It's one of the most strategically complex games in existence, with the number of board positions in the order of 10^170, dwarfing the estimated number of atoms in the observable universe, which is closer to 10^80.

Page had played Go with his Google cofounder Sergey Brin when they were building the company at Stanford years before, and when he mentioned his interest in the game to Hassabis a few weeks after the acquisition, Hassabis said his team could build an AI system that could beat a human champion.

Hassabis didn't just want to impress his new boss. As well as being an accomplished scientist, he was an exceptional marketer. He understood that if *AlphaGo* could beat a global champion of Go in the same way IBM's Deep Blue computer had beaten chess's Garry Kasparov in 1997, it would create a thrilling new milestone for AI and cement DeepMind's credibility as a leader in the field. DeepMind had its sights on South Korea's Lee Sedol and challenged him to a five-game match in Seoul in March 2016.

More than two hundred million people tuned in online and on TV to watch Lee play five games of Go against DeepMind's computer. The DeepMind scientist operating the program stopped drinking hours before the match so he wouldn't need a toilet break. Hassabis paced between *AlphaGo*'s control room and a private viewing area as it happened. He couldn't eat. His team had taught *AlphaGo*'s neural network thirty million possible moves.

To win Go, players need to capture their opponent's stones by completely surrounding them, and doing that requires various nuances in strategy: balancing the need to attack and defend, long-term versus short-term goals, and predicting the sequences of moves your opponent might make. That means carefully choosing on which lines of the grid you place your stones. The first lines closest to the edge are rarely used because they don't offer much chance of surrounding an opponent to capture territory, for instance. That's why in its second match against Lee, *AlphaGo* made what seemed like a bizarre mistake for its thirty-seventh move of the game. It played its stone on the fifth line from the right of the board. Typically, moves on the fifth line are seen as less effective because they give the opponent a territorial advantage on the fourth line. Playing on the fifth was considered wasteful. The move was so out of the blue and unconventional that Lee took fifteen minutes to consider his response and even walked out of the room.

"That's a very surprising move," said one commentator, who believed *AlphaGo*'s human operator had mistakenly clicked the wrong box on the board.

But about one hundred moves later, the odd strategy began to make sense. Two of *AlphaGo*'s black stones on the bottom left of the board ended up spilling over the other side and connecting up perfectly with the stone it had put on that fifth line. After four more hours of play, Lee resigned. He and the commentators went on to describe the thirty-seventh move as "beautiful." Hassabis said it showed glimmers of creativity in AI. All told, *AlphaGo* won four of its five matches against Lee.

It was a landmark moment for AI that gave DeepMind the biggest period of press attention it had ever received, including an award-winning Netflix documentary about *AlphaGo*. Hassabis was ready to end things on a high and retire the program so he could move on to the next project.

But Google also saw an opportunity. It wanted to showcase Google's technological prowess to Beijing and forge a new path back into China. *AlphaGo* could represent a new kind of Ping-Pong diplomacy with China, the executives thought, like the exchange of table tennis players between the United States and China in 1971 that helped thaw diplomatic relations after the Cold War. If the match in Korea had been a publicity stunt for DeepMind, the next one in China should be for Google.

Google wanted DeepMind to put *AlphaGo* in front of an even more advanced player, Ke Jie, a nineteen-year-old ranked as the world's number-one Go player at the time and who was based in China. Ke Jie was a completely different player than Lee Sedol, eager to taunt his opponents and puff himself up. But Google was being no less arrogant, believing it could flaunt its technology and win its way back into China.

The situation worried Hassabis, according to former DeepMind staff. If *AlphaGo* won, it would look like the big bad AI was out to beat humans again and again. If they lost, then all the hype they'd generated in South Korea would be wiped out. It seemed like a lost cause either way.

Knowing that Google desperately wanted that foothold in China, though, Hassabis used his strategic prowess to work out a compromise with Pichai: they'd do another match, but this time they'd use a new version of *AlphaGo* called *AlphaGo* Master. Instead of running on hundreds of different computers, it would run on just one machine powered by a Google chip. This way, they could frame the match as a test of their new AI system rather than another attempt to crush human champions. If the system lost, they could save face by saying it wasn't comparable to the original *AlphaGo*, but if it won, they could herald a new, more powerful system. Google could pitch its new machine learning platform called TensorFlow and get some big corporate clients in China to pay for its small but growing cloud computing business. Pichai agreed.

The match took place in Wuzhen, China, in May 2017, and while Google executives had spent the past year lobbying Chinese government officials to broadcast it across

China's TV and internet services, the match ended up being blocked for most of the country. The new *AlphaGo* won all three games against Ke Jie, and hardly anyone in China knew.

Google's leadership tried to stay positive about the situation. As he was being interviewed on stage at the match event, Schmidt used the opportunity to praise TensorFlow, saying that top Chinese internet companies like Alibaba, Baidu, and Tencent should try it. "All of them would be better off if they used TensorFlow," he said. Behind the scenes, Google was so desperate to get back into the Chinese market that it also reversed some of its previous resistance to Beijing's demands on censorship and even surveillance. According to a memo that was leaked to *The Intercept* in 2018, Google executives had ordered its engineers to work on a prototype search engine for China codenamed Dragonfly, which blacklisted certain search terms and linked people's searches to their mobile numbers. It was backtracking on its principles to help an oppressive regime surveil its citizens.

But Google's hunger for new business blinded it to the folly of trying to reenter China. Chinese technology firms were making big strides on AI research. They didn't really need TensorFlow—or Google, for that matter. The Chinese internet giant Baidu had even poached Andrew Ng, the Stanford professor who'd started Google Brain, from Google a year earlier. The Chinese government calculated that its citizens and its burgeoning tech sector could live without the search giant's services.

Two months after the Ke Jie match, Beijing revealed its latest long-term goal for the country, this time to become a world leader in artificial intelligence, surpassing the United States, by 2030. The government would fund a range of AI start-ups and moonshots, which collectively looked like its own version of the Apollo program. There was no mention of working with Google or any other Silicon Valley technology companies to do that.

It soon dawned on Google's executives that their dreams of entering the enormous Chinese internet market and watching its profits balloon were unrealistic. It was a huge disappointment for the company. Hassabis had also put himself in an awkward position with the success of *AlphaGo*. By creating a storm of positive publicity for DeepMind and showcasing its advanced AI, he'd made the lab look even more useful to Alphabet. Even so, Hassabis forged ahead with the plans he'd laid mostly with Suleyman to break away.

He was so confident it would happen that a few weeks after the May 2017 China match, Hassabis flew most of DeepMind's more than three hundred staff members up to a rural part of Scotland for a retreat, where he and Suleyman told all of them about the breakaway plan. At a hotel and conference center that they had rented out, they announced the plan to turn DeepMind into a separate global interest company. They told staff that DeepMind would eventually become a nonprofit organization in which Google would be a stakeholder and that it would be like other organizations with a public interest, like the United Nations and the Gates Foundation. The goal was to become an organization for good, they explained, and guide AI in a way that was positive for the world. Instead of being a financial asset of Google, DeepMind would enter an exclusive licensing agreement with the company instead, while pursuing its mission to solve the world's problems.

DeepMind's staff were thrilled by the idea, according to people who were there at the time. If you were an AI researcher, suddenly you had the best of both worlds. You were working for a tech company that offered a great salary and benefits, but you were also "solving intelligence, and then solving everything else." The founders said the separation would be finalized by September of that year, 2017.

Hassabis and Suleyman asked staff to keep the GIC project secret, which wasn't that unusual. Most DeepMinders had signed strict nondisclosure agreements that stopped them from talking about the company's plans and technology. But in this case, they were also being told not to talk about the spinout internally. Some were told to use code words, for instance, sometimes referring to the project as "watermelon," and they used encrypted messaging apps like Signal to talk about it. Some DeepMind leaders advised staff not to discuss it on corporate devices or apps like Gmail, according to former staff.

DeepMind's researchers believed the secrecy was due to concerns about what Google might do with AGI. As the year wore on, those suspicions gained some credibility when Google got involved with a military project. The US Department of Defense had launched what it called Project Maven in 2017 to try to use more AI and machine learning in its defense strategies, for instance by giving its drones computer vision to get better at targeting weapons. When Google got involved, it was expecting to make $250 million a year from the partnership, according to emails leaked to *The Intercept*. Massive internal protests prompted Google to shut the project down and decline to renew its contract with the Defense Department, and it validated DeepMind's worries about its AI being misused.

But progress on the spinout was slow. Hassabis and other executives would assure staff that the spin-off was "six months away," and then repeat the mantra several months later. After a while, the engineers started wondering if the plan was going to happen at all. It didn't help that its contours seemed hazy. Suleyman, for instance, told staff that he wanted DeepMind's new rules around working with Google to be legally enforceable, but he and other managers couldn't clarify how that would happen in practice. Suppose Google were to use DeepMind's AI for military purposes down the line. Could DeepMind sue Google? That wasn't clear. DeepMind staff were told to draw up guidelines that banned its AI from being used for human rights violations and "overall harm." But what did "overall harm" actually mean? Nobody knew.

Part of the problem was that DeepMind hadn't hired enough people to help it answer those questions. It had been bulking up on scientists and programmers to improve the performance of its AI models, but only a handful of staff were researching ethical ways to design AI. In 2020, for instance, most of DeepMind's roughly one thousand staff members were made up of research scientists and engineers, while fewer than a dozen were researching ethics and just two were working at a PhD level doing academic research on the issue. Almost no one was looking into how AI systems could lead to bias and racism or hurt human rights. “You can’t say you have an ethics team when we’re really only talking about two people,” one DeepMind staffer said at the time.

In AI, “ethics” and “safety” can refer to different research goals, and in recent years, their proponents have been at odds with one another. Researchers who say they work in AI safety tend to swim in the same waters as Yudkowsky and Jaan Tallinn and want to ensure that a superintelligent AGI system won’t cause catastrophic harm to people in the future, for instance by using drug discovery to build chemical weapons and wiping them out or by spreading misinformation across the internet to completely destabilize society.

Ethics research, on the other hand, focuses more on shaping how AI systems are designed and used today. They study how the technology might already be harming people. This is because the Google Photos algorithm that had labeled Black people as “gorillas” wasn’t an isolated example. Bias is an immense problem in AI. Algorithms used in the American criminal justice system have disproportionately, and incorrectly, flagged Black individuals as more likely to reoffend. And developers have used AI tools for ethically repulsive purposes, like the Stanford researchers who released a facial recognition system that claimed to distinguish people’s sexual orientation.

The people who built all three systems should have put more thought into designing their models with fairness, transparency, and human rights in mind. But these issues are squishy and hard to define, and they also tend not to impact the people running AI companies, who, more often than not, skew toward male and Caucasian. When AI systems go awry today, they are more likely to harm people of color, women, and other people from minority groups.

What’s perplexing is that in 2017, DeepMind was talking to the press and on its website about the importance of ethics in its “mission of solving intelligence to advance science and benefit humanity.” It spoke to *Wired* magazine, for instance, about how its small team of ethics researchers would swell to twenty-five people within the next year.

But in reality, that team grew to only about fifteen, in large part because DeepMind’s leaders were so focused on the spin-out project, according to a former executive. “They talk about this work all the time but it’s just a handful of [ethics researchers], struggling,” another staffer said, explaining that the ethics team had no support team and few resources. “It doesn’t make sense. This is a [multibillion-dollar] corporation.”

If DeepMind wasn’t putting its money where its mouth was on ethics, that raised questions about why the founders were so keen to spin out from Google in the first place. Did they really care about preventing their technology from doing harm, or were they feeding a more personal instinct to maintain control? As part of the terms of its separation, DeepMind was planning to sign an exclusive licensing agreement with Google, but the founders couldn’t seem to clarify where they drew the line on using their AI for weapons or whether that was legally enforceable. They seemed big on ambition but lacking on details. Some staff wondered if Hassabis, Suleyman, and Legg were being naive in wanting to have their cake and eat it—taking Google's money to keep building AGI but grabbing control away from Google too.

In much the same way Google had started with a "don't be evil" motto, DeepMind's founders had kicked off their life under Google with good intentions. They'd left $150 million on the table with Facebook to keep an ethics board. But years later, they seemed to be prioritizing performance and prestige over ethics and safety. They didn't have clear answers for how they would contain AGI other than hiring an all-star team of mathematicians like Terence Tao or on how they would stop that technology being used in harmful ways.

All of this raised a bigger question. Could you even do meaningful work on ethical AI from inside a large corporation? The answer came from inside Google itself. It was a resounding no.

CHAPTER 8

Everything Is Awesome

To understand why it became so maddeningly difficult to design ethical AI systems at Google, or to even turn innovative ideas into products at the company, you have to step back and look at some numbers. At the time of writing, Google's parent company, Alphabet Inc., had a market capitalization of $1.8 trillion. In 2020, Apple became the first publicly traded US company to hit a $2 trillion valuation, while Amazon's and Microsoft's market values were hovering at around $1.7 trillion and an astonishing $3 trillion, respectively. Before Apple first became a trillion-dollar company in 2018, no company ever had become so big. Yet there's one thing that nearly all the world's most valuable companies have in common: they are tech firms. In fact, the companies that we might normally think of as being gargantuan are only a quarter the size of their peers in Silicon Valley. Oil giant Exxon Mobil is valued at a piddling $450 billion while Walmart is worth $435 billion. Combine the market caps of the tech giants and you have surpassed the gross domestic product of most of the world's nations, barring the US and China.

Looking back in history, the companies that we once thought of as giants also pale in comparison to those of today. At its peak before being broken up in 1984, AT&T had a market capitalization of around $60 billion in 1984 dollars, or about $150 billion in today's money. General Electric's highest market cap was about $600 billion in 2000.

Even the market dominance of tech giants is unparalleled. Before regulators broke it up in 1911, Standard Oil controlled about 90 percent of oil business in the United States. Today, Google controls about 92 percent of the search engine market—globally. Roughly one billion people around the world run a search on Google each day. More than two billion check Facebook. And about 1.5 billion people in the world have an iPhone. No government or empire in history has touched so many people at once.

It's taken a little over two decades for these companies to reach this scale since the dotcom boom and bust. How did they get so big? They bought companies like DeepMind, YouTube, and Instagram, and they sucked up a prodigious amount of data about consumers, allowing some of them to target us with advertisements and recommendations that could influence human behavior on a massive scale. While Google collects data through search queries and YouTube interactions, Amazon tracks our purchases and browsing behavior. The sheer scale of data they collect is hard for regular people to fathom, including personal details, browsing history, location data, and even voice recordings in some cases. The data isn't just voluminous but diverse, giving tech companies a detailed picture of consumer behavior.

Firms like Facebook and Google use that data to conduct hypertargeted advertising, displaying ads that pique a person's interests and fuel sophisticated recommendation algorithms. That software powers the "feeds" that people thumb through every day, making sure the content that pops up is most likely to keep them continuously scrolling. The companies are incentivized to keep us as addicted as possible to their platforms, since that generates more ad dollars. But the adverse effects are plentiful. Americans are so addicted to Facebook, Instagram, and other social media apps that they checked their phones 144 times a day on average in 2023, according to one study.

All that personalized "content delivery" has also amped up the generational and political divisions between millions of people, since the most engaging content tends to be the kind that provokes outrage. Facebook, for instance, often recommended the most provocative political content in people's feeds during the 2016 US presidential election and exposed many users to news and opinions that reinforced their existing beliefs, creating echo chambers. The same phenomenon fueled growing resentment toward immigration in Britain in the months leading up to the UK's referendum on Brexit, as well as hostility toward the Rohingya people of Myanmar in 2017. Facebook's algorithms supercharged the spread of hateful content against the Rohingya so much that it fueled the Myanmar military in its genocidal campaign to kill, torture, rape, and displace the Muslim ethnic group in the thousands, according to a report by Amnesty International. Facebook has admitted in press reports that it didn't do enough to prevent the incitement of violence against the Rohingya.

For all the division Facebook sowed, the company's business model was unfathomably successful, treating its billions of users and their data as the product and its advertisers as the real customers. The more data it could get, the more it could make from advertisers. While this engagement-based model had toxic effects on society, it incentivized Facebook to do one thing: become as big as possible.

The other way these companies became so enormous was network effects, a seemingly magical phenomenon that every start-up founder craves. The basic idea of network effects is that the more users and customers a company has, the better their algorithms will become, making it increasingly difficult for competitors to catch up, further entrenching their grip on the market. In the case of Facebook, for instance, people started joining the site because everyone else was on Facebook, and many have stayed on the site for years since—or at least resisted the urge to delete their accounts—for much the same reason. If you're an Apple fangirl or fanboy, you'll know how hard it is to try another device maker like Samsung, or to get the latter's accessories to work with an iPhone. All these interconnected products and services make it hard to switch, reinforcing Apple's dominance.

We have no historical reference point for what happens when companies become this big. The market cap numbers that Google, Amazon, and Microsoft are currently achieving have never been seen before. And while they bring greater wealth to the shareholders of those companies, including pension funds, they have also centralized power in such a way that the privacy, identity, public discourse, and increasingly the job prospects of billions of people are beholden to a handful of large firms, run by a handful of unfathomably wealthy people.

It is little wonder that for those working inside a tech giant who see something wrong, sounding the alarm can seem as futile as trying to turn the *Titanic* around just moments before hitting the iceberg. Still, that didn't stop an AI scientist named Timnit Gebru from trying.

In December 2015, at the NeurIPS conference where Sam Altman and Elon Musk announced they were creating AI "for the benefit of humanity," Gebru looked around at the thousands of other attendees and shuddered. Almost no one there looked like her. Gebru was in her early thirties and Black, and she'd had anything but a conventional upbringing with the support system that many of her peers had enjoyed.

Her Eritrean father, an electrical engineer, died when she was five, and she fled war-torn Ethiopia when she was a teen. Her teachers at high school in Massachusetts took a dim view of her ambitions as a new immigrant. They discouraged her from taking Advanced Placement courses, saying she might find them too difficult. She recalled one teacher telling her, "I've met so many people like you who think that they can just come here from other countries and take the hardest classes," according to a profile of Gebru in *Wired* magazine. But Gebru took them anyway and won a place studying electrical engineering at Stanford University.

Eventually, she stumbled onto the field of artificial intelligence and computer vision, software that could "see" and analyze the real world. The technology was entrancing, but Gebru saw red flags. AI systems were being given powerful roles in people's lives, from giving someone a credit score to granting a mortgage, and from flagging someone's face for the police to helping a human judge decide a criminal sentence. While it seemed like these systems could be the perfect neutral arbiter, they often were not. If the data they were trained on was biased, so was the system. And Gebru was painfully aware of bias.

When she was a young woman in San Francisco, for instance, she and another Black woman were attacked and strangled by some men at a bar. They went for help, but the police accused them of lying and detained them in a cell. Another time, while Gebru was writing her thesis at Stanford, she learned that only one other Black person had done a computer science PhD there. And out of the five thousand people at the biggest international AI conference where Altman and Musk were launching OpenAI in 2015, only five participants were Black.

Gebru knew these weren't one-offs. Bias was systemic in the world around her. Decades after the success of the twentieth century's civil rights era, racism was still culturally baked into the world's institutions and psyche. AI could make that worse. For a start, it was typically designed by people who hadn't experienced racism, which was one reason why the data being used to train AI models also often failed to fairly represent people from minority groups and women.

Gebru saw the consequences of that in her academic research. She came across an investigation into software being used in the US criminal justice system called COMPAS (Correctional Offender Management Profiling for Alternative Sanctions), which judges and parole offices used to help make decisions about bail, sentencing, and parole.

COMPAS used machine learning to give risk scores to defendants. The higher the score, the more likely they were to reoffend. The tool gave high scores to Black defendants far more than white defendants, but its predictions were often erroneous. COMPAS turned out to be twice as likely to be wrong about future criminal behavior by Black defendants as it was for Caucasian ones, according to a 2016 investigation by ProPublica, which looked at seven thousand risk scores given to people arrested in Florida and checked if they'd been charged with new offenses in the next two years. The tool was also more likely to misjudge white defendants who went on to commit other crimes as low-risk. America's criminal justice system was already skewed against Black people, and that bias looked set to continue with the use of inscrutable AI tools.

While writing her PhD thesis at Stanford, Gebru pointed to another example of how authorities could use AI in disturbing ways. She trained a computer vision model to identify twenty-two million cars shown on Google Street View, then dug into what those cars might say about an area's demographics. When she correlated the cars with census and crime data, she found that areas with more Volkswagens and pickup trucks tended to have more white residents, while those with Oldsmobiles and Buicks had more Black ones. And areas that had more vans also had more reports of crime. Such correlations could be exploited. What if police used that data to try to predict where crimes might take place, like in the movie *Minority Report*?

The idea wasn't crazy. For several years, police precincts across the US had been using computers to advise their officers on what areas to patrol, a technology known as predictive policing. But the software was trained on historic data, which meant it often led them to target minority communities. If the data showed a community was being overpoliced, the software would lead that community to continue being overpoliced, amplifying a pre-existing problem.

AI was spreading other stereotypes online, too, in subtle but insidious ways. Both Google Translate and Microsoft's Bing Translate sometimes made certain professions male when translating them to other languages. The phrase *o bir muhendis* in Turkish, which has gender-neutral pronouns, became *he is an engineer* in English, while *o bir hemsire* became *she is a nurse*. Software made these assumptions thanks to a popular technique called word embedding, which looked at words that tended to hang around other words, like "engineer." The model then figured out what other word fit best, like "he." Google, Facebook, Netflix, and Spotify all powered their online recommendations with the word embedding technique, in spite of the true-to-life gender imbalances they were introducing into their software.

Clearly AI had problems that should have been dealt with yesterday, so when Sam Altman announced OpenAI in 2015, Gebru was livid. She started writing an open letter about how wasteful it was for a few egocentric billionaires like Musk and Thiel to put their money into an effort to build godlike AI, and complained that the only concerns raised about the new nonprofit had been that its researchers were too focused on deep learning.

"A white tech tycoon born and raised in South Africa during apartheid, along with an all-white, all-male set of investors and researchers is trying to stop AI from 'taking over the world' and the only potential problem we see is that 'all the researchers are working on deep learning?'" she wrote. "Google recently came out with a computer vision algorithm that classified Black people as Apes. AS APES. Some try to explain away this mishap by stating that the algorithm must have picked out color as an essential discriminator in classifying humans. If there was even one Black person [on] the team, or just someone who thinks about race, a product classifying Black people as apes would not have been released.… Imagine an algorithm that regularly classifies white people as nonhuman. No American company would call this a production-ready person detection system."

One of Gebru's colleagues told her not to publish the letter. It was too candid, and she'd probably be identified. Gebru decided to keep it private (only releasing it a few years later), but she couldn't help asking why some of Silicon Valley's most powerful people were so worried about the possibility of AI doom when AI was already causing real harm to people today. There were two answers. The first was that few if any of the leaders of OpenAI and DeepMind had ever been, or ever would be, on the receiving end of racial or gender discrimination. The second was that it was, paradoxically, in their corporate interests to shout about the risks of an all-powerful superintelligence. It might not make perfect sense to warn people about the dangers of something you are trying to sell, but it was a brilliant marketing strategy. People tended to care more about the here and now than their long-term future. If AI suddenly looked like it was on track to snuff us out down the line, that also gave it the alluring glow of impressive capabilities today.

The strategy was also a clever way to avert the public's attention from the thorny, more immediate problems that companies could take action on, requiring them to slow down their development and rein in the capabilities of their AI models. One way to limit AI models from making biased decisions was to spend more time analyzing the data they were trained on. Another was to make them narrower in scope, which would blow a hole in the goal of giving AI systems the power to generalize their knowledge.

This wouldn't be the first time large companies had distracted the public while their businesses swelled. In the early 1970s, the plastic industry, backed by oil companies, began to promote the idea of recycling as a solution to the growing problem of plastic waste. Keep America Beautiful, for instance, was an organization founded in 1953 that ran public service campaigns encouraging consumers to recycle, and was funded in part by drinks and packaging firms. Its famous “Crying Indian” ad aired on Earth Day in 1971 and encouraged people to recycle their bottles and newspapers to help prevent pollution. If they didn’t, they were guilty of showing flagrant disregard for the environment.

Recycling is not a bad thing per se. But by promoting the practice, the industry could argue that plastics weren’t inherently bad so long as they were recycled properly, which shifted the perception of responsibility from producers to consumers. Plastics companies knew that recycling on a large scale was expensive and often inefficient. A 2020 investigation by NPR and the PBS series *Frontline* found that less than 10 percent of plastic had ever been recycled, despite decades of public awareness campaigns.

What the campaigns had achieved, though, was drawing public attention away from questioning the rapid expansion of plastic production and the toll that was having on the environment. Recycling became part of the public discourse. News publications, consumers, and policymakers in Washington spent more time talking about how to do more recycling than they did about regulating actual plastic production by companies.

In just the same way Big Oil redirected the world’s attention from their own significant environmental impact, AI’s leading builders could exploit the buzz around a future Terminator or Skynet to distract from the present-day problems that machine learning algorithms were causing. The burden of responsibility wasn’t on the creators or the industry to act now. It was an abstract problem to be dealt with later.

In January 2017, a few months before DeepMind tried to help shoehorn Google back into China with *AlphaGo*, Gebru presented the findings of her thesis to an audience of venture capitalists and executives from Silicon Valley. As she clicked through her slides, she explained that AI systems could combine their ability to recognize cars with their ability to make predictions to forecast things like voting patterns or household income.

One venture capitalist there, a Tesla investor and friend of Elon Musk named Steve Jurvetson, was stunned, but not for the reasons Gebru was hoping. Think about how powerful this kind of data made Google and the kinds of insights it could make about different neighborhoods or towns. He was so impressed that he posted photos of Gebru’s talk to Facebook.

In what was a continuing incongruity in AI, some in the room saw a financial opportunity while others like Gebru saw a danger that needed to be contained. Each time AI’s capabilities grew, an unintended consequence arose that often caused harm to a minority group. Facial recognition systems were nearly perfect at recognizing the faces of white men, but often made mistakes with Black women. A 2018 landmark study from MIT graduate researcher Joy Buolamwini found that face-recognition systems from IBM and Microsoft and China’s Face++ were more likely to misclassify the gender of darker-skinned and female faces, something she noticed when her own face wasn’t recognized by a similar program. Many of these systems were being trained on datasets of photos that were dominated by Caucasian males and on photos scraped from the web. The databases overrepresented them because the internet reflected the demographic of Western people who had more access to it.

Gebru wasn't throwing her hands up in the air. She had solutions. One was for the creators of AI systems to follow more rigorous standards when training their models. After joining Microsoft, she wrote up a set of rules called "Datasheets for Datasets," which said that when training an AI model, programmers should create a datasheet that showed all the details about how it was created, what was in it, how it would be used, what its limitations might be, and any other ethical considerations. It promised to be an exasperating extra step in paperwork for AI builders, but it had a purpose. If the model ended up being biased, it'd be a lot easier to figure out why.

Figuring out why AI systems make mistakes is much harder than people think, especially as they become more sophisticated. In 2018, Amazon realized that an internal AI tool that it used to sift through job applications kept recommending more male candidates than female candidates. The reason: the tool's creators had trained it on résumés submitted to the company over the previous ten years, most of which came from men. The model had learned that résumés with male attributes were more desirable as a result. But Amazon didn't—or wasn't able to—fix the tool. It just shut it down completely.

Google took a similarly blunt approach when its Photos tool had labeled some Black individuals "gorillas," stopping the app from identifying gorillas completely even as it continued to recognize other animals. The original, painful mistake came about because Google hadn't trained its tool on enough images of Black and dark-skinned people and likely hadn't tested it on employees enough either. But even in late 2023, the company still wasn't confident enough in its ability to tweak its AI model to fix it, so it simply shut the feature down.

Some AI researchers say it's too difficult to fix these biases, arguing that modern-day AI models are so complex that even their creators don't understand why they make certain decisions. Deep-learning models, like neural networks, are made up of millions or billions of parameters, also known as "weights," that act as adjusters in complex mathematical functions between connected layers. Think of the layers of a neural network as being a bit like a factory with an assembly line, where each person on the line has a certain job like painting a toy car or adding the wheels. By the end of the line, you have a toy car. Each layer in a neural network is like the stations in the assembly line, making their own little adjustment to the data. The problem is that with so many small changes happening in sequence, it's difficult to trace back exactly what each station on the assembly line (or layer in the neural network) did to make the toy car—to make the decision to label a Black defendant as being at high risk of reoffending.

After Google was in the news for the gorilla debacle, another computer scientist, named Margaret Mitchell, joined the search giant to try to prevent similar mistakes from happening. Born in Los Angeles and well-known among AI researchers for her work on fairness in machine learning, Mitchell joined a small but growing effort in the field to be more careful about the real-world impact of machine learning systems. Like Gebru, she was worried about the strange mistakes AI systems were making. She'd conducted most of her postgraduate research in computational linguistics and then natural language generation, studying all the ways that computers could describe objects or analyze emotions in text. When she worked on an app for blind people at Microsoft, she was unnerved when it described a Caucasian person like herself as a "person" but someone with dark skin as a "Black person."

Another time, she was running some experiments on a neural network that described images, and she fed it some pictures of a factory explosion in England. One photo had been taken from high up in a nearby apartment. It showed plumes of smoke and a TV news channel in the foreground reporting on the incident. Mitchell was stunned when the AI system told her the image was "awesome," "beautiful," and a "great view."

"The system had an everything-is-awesome problem," says Mitchell, recalling the famous song from *The Lego Movie* about a brick-based world where all of life's troubles are brushed under the carpet. "It had no concept of mortality, and it had no concept that death is bad."

What it had really learned from the training photos was that sunsets were beautiful and that being high up gave you a great view. That's when the penny dropped for Mitchell. Data was everything. By creating gaps in the data to train her own system, she had encoded it with all kinds of biases, including ones that dismissed the loss of human life.

As Mitchell worked on these issues at Google, she noticed another frustrating feature about working for a large tech firm. She was in the middle of a stifling bureaucracy, chasing endless meetings and managers who seemed perpetually worried about the company's reputation.

In 2018, Mitchell sent Gebru an email asking to join her at Google. AI ethics was small enough that the two already knew each other. Would Gebru co-lead Google's ethical AI research team?

Gebru hesitated. She'd heard rumors through the grapevine that Google was a toxic place to work, particularly for women and minorities. There was no better example than the case of Google executive Andy Rubin. Rubin had been a rock star at Google, having cofounded its popular Android operating system, but in 2014, he quietly left the company following allegations of sexual misconduct. A few years later, an investigation by the *New York Times* found that Google's management had looked into the sexual misconduct allegations and found them to be credible. Yet instead of kicking Rubin out the door, Google had given him a hero's goodbye, which included a $90 million exit package.

It wasn't all bad at Google, though. Gebru was impressed by how staffers would stand up and fight back when they saw the company do wrong. Thousands had staged a global walkout over Rubin's golden parachute, and months before she joined, more than three thousand employees had signed an open letter to CEO Sundar Pichai, demanding that the company withdraw from Project Maven—which it did. Even better, those protests had been coordinated by an expert in AI ethics, a woman named Meredith Whittaker, whose clear articulation of the problem forced Google to reconsider the program. Maybe this was a place where she could promote more responsible practices like her Datasheets for Datasets standards.

But looking at the size of her new ethics team, it was clear where tech giants like Google prioritized their investment in AI: capabilities. Despite the importance of their work, the team was just a handful of computer scientists. Across the rest of the company, thousands of engineers and researchers were still working on making the company's AI systems faster and better, creating new standards of capability that Gebru and Mitchell were constantly chasing and trying to scrutinize for unintended consequences.

Mitchell was feeling ground down at Google. When she warned managers at meetings about some of the potential problems their AI systems could introduce, she'd get emails from the Human Resources Department telling her to be more collaborative. In Silicon Valley, women made up only about a quarter of computing jobs at firms like Google, Apple, and Facebook and were still earning eighty-six cents for every dollar earned by men in 2020. Women often experienced unequal treatment, harassment, and discrimination in hiring and promotions, and the situation was especially challenging for Black women. Many of the women who showed up to a typical Silicon Valley conference or drinks event worked in marketing or PR rather than in engineering or research. Women were thus more likely to work on AI ethics in the first place; they knew firsthand what discrimination could feel like. But it also meant they struggled to be the loudest voices in the room.

Still, Mitchell found herself surprised and then in awe of Gebru, who was audacious and had no qualms about standing up to authority if she needed resources or saw wrongdoing. One day when the pair were sitting in Gebru's office at Building 41 of the Google campus, they were talking about an upsetting email that had come through from one of their managers, which reflected the discrimination they both felt at the company. Mitchell was on the verge of tears. Gebru took a different view.

"Don't be depressed," she told Mitchell. "Get angry."

Gebru pulled her laptop toward her and started drafting a response to the manager, reading it aloud as she went along and clinically taking apart their manager's points. Later, when both Mitchell and Gebru would be fired by Google, that same manager would publicly vouch for them both and then resign shortly after.

Gebru and Mitchell were about to finally bring proper attention to their cause, even if it meant being kicked out in what would become a public scandal. But they were racing against Google's central effort. That much bigger team of scientists who were tasked with making Google’s AI smarter were about to stumble upon one of the biggest leaps forward in AI history. It was a miracle they did.

CHAPTER 9

The Goliath Paradox

In 2017, Google had about eighty thousand salaried employees. Not all of them were engineers. There were curators of the daily Google Doodle that showed up above everybody's search bar. There were in-office chiropractors and masseuse managers, snackologists who made sure the staff were fueled between their three hot meals at the canteen, horticulturists who looked after the plants, and cleaners who wiped down the foosball tables.

Google's business model was a golden goose. That year its advertising business was generating close to $100 billion annually—a number that would more than double by 2024—so it was natural that much of that money would go toward bulking up its talent. Silicon Valley tended to measure success with two metrics: how much money you had raised from investors, and how many people you had hired. Gargantuan employee numbers reflected the empire-building dreams of CEOs like Larry Page and Sergey Brin, even if it wasn't always clear what many of their middle managers were doing.

Google's corporate bloat was not unusual. Facebook at that time had about 40,000 employees and Microsoft had 124,000, while start-up founders dreamed of running their own corporate campuses complete with gyms and free ice cream stalls. Demis Hassabis was one exception to the rule, perhaps because he was stationed an ocean away. He didn't want DeepMind to get sucked into the distracting world of Silicon Valley perks and size obsessions.

The problem with being so big was that if someone did invent something groundbreaking inside Google, it might struggle to see the light of day. Google's digital ad business was sacrosanct. You didn't mess with the algorithms that powered it unless you really had to. For all the kudos that Silicon Valley got for being the innovation capital of the world, its biggest companies weren't all that innovative. Google's home page had barely changed over the past decade. The iPhone was still the same old rectangular slab of metal. And nearly every new Facebook feature was a direct copy of a competitor like Snapchat or TikTok. Once these companies reached a stage where their revenue was in the tens of billions, messing with their formula for success was too dangerous.

That's why when a group of researchers at Google made one of the most important discoveries in artificial intelligence in the past decade, the search company left it to languish internally. Their story in a nutshell showed how the monopolistic scale of Big

Tech firms crimped their ability to invent, forcing them to react to other people's innovations by copying and buying them outright. But this particular negligence was worse for Google. In the end, OpenAI not only capitalized on Google's big invention, it used that invention to launch the first viable threat to the search giant in years.

The *T* in ChatGPT stands for "transformer." This has nothing to do with the alien robots that morph into eighteen-wheelers but a system that allows machines to generate humanlike text. The transformer has become critical to the new wave of *generative* AI that can produce realistic text, images, videos, DNA sequences, and many other kinds of data. The transformer's invention in 2017 was about as impactful to the field of AI as the advent of smartphones was for consumers. Before smartphones, mobile phones couldn't do much more than make calls, send texts, and play the odd game of Snake. But when touch-screen smartphones hit the market, suddenly people could browse the internet, use GPS, take high-quality photos, and use millions of different apps.

Transformers also broadened the scope of what AI engineers could do. They could handle far more data and process human language much more quickly. Before transformers, talking to a chatbot felt like talking to a dumb machine because the older systems operated on sets of rules and decision trees. If you asked a bot something that wasn't already written into its programming (which was often likely), it would be stumped or make a peculiar error. That was how digital assistants like Apple's Siri, Amazon's Alexa, and even Google's Assistant were originally designed. They treated each query as a single isolated request, which meant they were terrible at context. They couldn't remember questions you'd previously asked in the same way a person would during a conversation. For instance:

> "Alexa, what's the weather like in Indianapolis right now?"
>
> "Right now in Indianapolis, it's twenty-four degrees Fahrenheit with cloudy skies."
>
> "How many hours would it take for me to fly there from London?"
>
> "From London, to fly to your current location, it would take about forty-five minutes."

My current location at the time of writing was Surrey, which would presumably take forty-five minutes to fly to from London's Heathrow Airport. How Alexa came up with that convoluted flight plan doesn't matter—the problem was that it couldn't process that "there" meant Indianapolis, which I'd asked about two seconds earlier. The systems behind most of these traditional digital assistants were narrow and still relied primarily on keywords. That's why they still gave canned responses.

Transformers broke chatbots out of those bonds. They could deal with nuance and slang. They could refer back to that thing you said a few sentences earlier. They could handle almost any random query and give a personalized answer. One word summed up the upgrade: they were more *general*. And for many AI researchers, that meant a step toward AGI. It would also open a debate about whether computers were starting to "understand" language in the same way humans did or if they were still just processing it through math-based predictions.

In one way it's astonishing that the invention came out of Google at all. For all the talent and resources the company had, its bloat and fear of disrupting its ad business hampered staff who tried to push through new innovations. Google Brain had the company's most advanced deep-learning researchers, but they grappled with unclear goals and strategies from management, former staff there say. The culture of complacency partly came from having so many talented scientists on staff, like Geoffrey Hinton. The bar was high, and Google was already using cutting-edge AI techniques, like recurrent neural networks, to process billions of words of text every day.

If you were a young AI researcher like Illia Polosukhin, you were sitting next to the people who'd invented these techniques. In early 2017, Polosukhin was getting ready to leave Google and was willing to take some risks. Inside one of the Google canteens, two floors below the office of Larry Page, the twenty-five-year-old Ukrainian was spitballing with two other researchers, Ashish Vaswani and Jakob Uszkoreit. His lunch mates also didn't like following the conventions of other scientists in the building. Vaswani was hungry to work on a big project. Uszkoreit had been at Google for more than ten years and was wary of the way Google Brain's incentive structure had morphed into something like a glorified academic institution; after hiring dozens of new graduates and academics, he was surrounded by people who mainly cared about being first author on a paper or getting published at a conference. Whatever happened to making great products?

Uszkoreit would get impressed looks from people at a party if he mentioned where he worked. But whenever he added that he worked on Google Translate, they'd start laughing. The service was clunky and often inaccurate, particularly on non-Latin languages like Chinese. Polosukhin agreed that Google Translate sucked. He had friends in China who complained about the service. Uszkoreit wondered aloud if there was a better way. Google engineers tended to believe they were already working with the most advanced technology, so they operated on a motto of "if it ain't broke, don't fix it." Uszkoreit saw things a different way: if it ain't broke, break it.

"What if we got rid of the recurrent neural networks in the decoder of machine translation and just used attention in the decoder," one of them asked. "Wouldn't that speed up inference time?"

In AI-speak, the researchers were asking if they could take better advantage of ultrapowerful computing chips. Till then, Google had been using a technique called recurrent neural networks to analyze words. The system looked at each word in a sequence, as you might when reading a sentence from left to right. This was the cutting-edge method at the time, but it didn't take full advantage of the hefty chips that companies like Nvidia were making that could process lots of tasks at the same time. The chip in your home laptop probably had something like four "cores" to handle instructions, but the GPU chips used in servers to process AI systems had thousands of cores. This meant an AI model could "read" lots of words in a sentence all at once, not just in sequence. Not capitalizing on those chips was like switching off an electric saw to manually cut wood. Imagine unplugging an entire sawing machine and dragging its blade across a plank over and over. That would be a slow and arduous waste of the machine's potential. The same was happening with the AI systems that processed language. They weren't using the full potential of the chips that powered them.

Researchers like Vaswani had been looking into the concept of "attention" in AI, which is when a computer can pick out the most important information in a dataset. Over their salads and sandwiches, the trio wondered if they could use that same technique to translate words more quickly and accurately.

Over the next few months, the researchers started running experiments. Uszkoreit would scribble diagrams of the new architecture on white boards around the office, which passersby would look at with quiet skepticism. What his team were working on didn't make sense at the time. They were talking about removing the "recurrent" element of recurrent neural networks, which was crazy. And the different architectures that Vaswani was building still weren't significantly better than the status quo. But as word got around about their project, others wanted in.

One of them was Noam Shazeer, and he was already a legend at Google. He had coinvented a system that helped Google's Adsense program figure out which ads to show on which web pages. With his perpetually wide grin and booming voice, he was seen as quirky and chatted to suits like Sundar Pichai as if they were old friends. Shazeer had extensive experience with large language models. These were computer programs that could analyze and generate humanlike text after being trained on billions of words. Soon after he joined the ragtag group of researchers, Shazeer figured out some tricks that helped the new model work with large amounts of data.

"Once you put those things together, magic happened," Uszkoreit remembers. "That's when a whole bunch of other ideas took off."

Soon there were eight researchers working on the as-yet-unnamed project, writing code and refining the architecture of what they were calling the transformer. The name referred to a system that could transform any input into any output, and while the scientists focused on translating language, their system would eventually do far more.

After a while, they started to notice some improvements. "Oh, wow, this is different," Uszkoreit said at one point. The system was churning out long, complex sentence structures in German, and as someone who was fluent in the language, having spent many years in Germany as a child, Uszkoreit noticed it was better than the usual content Google Translate spat out. It was fluent, readable, and, most importantly, factually correct. Polosukhin, who spoke French, was noticing the same thing.

Llion Jones, a Welsh programmer on the team, was stunned to find that the system was doing something called coreference resolution. This had been a huge sticking point in the effort to make computers process language properly. It referred to the task of finding all expressions that refer to the same entity in a text.

For instance, in the sentence "The animal didn't cross the street because it was too tired," it's obvious to us as humans that *it* refers to the animal. But change the sentence to “The animal didn’t cross the street because it was too wide,” and *it* now refers to the road. Until then, it had been extremely difficult to get AI to infer that kind of shift in context because doing so required some element of commonsense knowledge, built up over years of experience of how the world works and how objects interact.

“It’s a classic intelligence test [that] AI’s failed on,” Jones says. “We couldn’t get common sense into a neural network.” But when they fed those same sentences into the transformer, the researchers could see something unusual happening to its “attention head.” The attention head was like a mini-detector in their model that focused on different parts of data it was being fed. It was the part that would harness the power of current chips and what allowed the transformer to pay attention to all the different words of a sentence at the same time, instead of one by one in sequence.

When the researchers changed the word from *tired* to *wide*, they could see the attention head switching *it* from being the animal to being the road.

“I don’t think anyone had seen that before,” Jones remembers. He almost started to wonder if he was getting a rare glimpse of real intelligence. “The fact that it was extracting common sense from unstructured text, that was evidence of something more interesting going on.”

About six months after those first conversations over lunch, the researchers wrote up their findings. Polosukhin had already left Google, but everyone else kept the project going and stayed in the office till midnight to wrap everything up. Vaswani, who was the lead author, slept on a nearby couch overnight.

“We need a title,” he said aloud at one point.

Jones looked up from his desk, nearby. “I’m not very good with titles,” he replied. “But how about ‘Attention is all you need’?” It was a random thought that had popped into his head, and Vaswani didn’t say anything in agreement. In fact, he got up and walked away, Jones recalls.

But later, the title “Attention Is All You Need” landed on the front page of their paper, a perfect summary of what they’d discovered. When you used a transformer, your AI system could *pay attention* to large amounts of data at the same time and do far more with it.

“I like to think of them as engines for reasoning,” says Vaswani.

These engines for reasoning had the potential to supercharge AI systems, but Google was slow off the mark to do anything about them. It took several years, for instance, for Google to plug transformers into services like Google Translate or BERT, a large language model that it developed to make its search engine better at processing the nuance of human language.

The transformer’s inventors couldn’t help but feel frustrated. Even a small start-up in Germany had started using the transformer to translate languages well before Google, putting the bigger company in a position where it was now playing catch-up.

Some of them tried to show Google the bigger possibilities of what the transformer was capable of. Not long after their paper was published, Shazeer started working with a colleague to use the technology on a new chatbot called Meena. They trained it on about forty billion words from social-media conversations on the public internet and eventually came to believe it would revolutionize how people searched the web and used computers. Meena was so sophisticated that it could improvise puns or banter with a human just as easily as it could hold a philosophical debate.

Shazeer and his colleague were excited about what they had just created, and they tried sending details of the bot to external researchers, hoping to launch a public demo and improve the clunky Google Assistant that people had in their homes in the form of a speaker with something much more sophisticated. But Google executives stopped those efforts. They worried that the bot would make outlandish remarks that hurt Google's reputation or, more specifically, its $100 billion digital advertising business. According to a report in the *Wall Street Journal*, they thwarted every attempt by Shazeer to launch Meena to the public or build it into Google products.

"Google doesn't move unless it's a billion-dollar business," says Polosukhin. "And it's really hard to build a billion-dollar business." That's why so many Google staff members had left to start two thousand different companies, according to an interview Pichai gave to *Bloomberg* in 2023. While that appears to frame Google as a fount of innovation, in reality the search giant was more like a giant squid sucking all the innovation out of the room. Many of those entrepreneurs who started new companies have also sold their firms back to Google or taken investment from the company. Where Google can't innovate, it usually buys.

There are two ways of looking at Google's sluggish approach to new technology. Publicly, it has framed itself as cautious. And many researchers at the company agree that its executives genuinely want to be careful about rolling out AI in a way that might harm society. In recent years, it drew up a list of guiding principles for using AI, largely copied from similar rules designed at DeepMind. In 2018, Google's top lawyer, Kent Walker, announced that the company would stop selling facial recognition technology because of the potential for abuse. And more broadly, Google has a process of putting its algorithms through a rigorous internal review process, sometimes including external peer review to scrutinize any ethical trade-offs.

But the company still makes ethically tone-deaf decisions. In May of that year, Pichai demonstrated a new assistant feature called Duplex, an AI voice that called a restaurant to book a table and used verbal tics like "um" and "uh" to make it sound eerily human. Pichai ended the demo to whoops of applause, but the service failed to disclose that it was a machine. Critics rounded on Google for deceiving humans on the other end of the line.

Google's cautious approach was largely a product of bloat. The downside to being one of the largest companies of all time, with a monopolistic grip on the search market, is that everything moves at a snail's pace. You're constantly afraid of public backlash or regulatory scrutiny. Your prime concern is maintaining growth and dominance. So intent has Google been on keeping a stranglehold on the search market that it paid more than

$26.3 billion in 2021 to Apple, Samsung, and others—more than a *third* of its net profit that year—just to preinstall its search engine on their phones, according to a recent landmark antitrust lawsuit brought by the US Department of Justice.

The company's sheer size and fixation on growth meant that its researchers or engineers often had to wade through several layers of management to sign off on even small ideas. And with virtually no competition, since Google controlled about 90 percent of all the world's online searches, there was no urgency to innovate.

At one point, back when the transformer group was pulling together its research, Shazeer had found himself chatting directly to Pichai next to one of the company's many coffee machines. His legacy of being an AI maven at Google for years had engendered personal connections with some of the company's top leaders. "This will replace Google entirely," Shazeer boasted of the new invention, according to one of his coauthors on the transformer paper, Lukasz Kaiser, who was there.

"He had this sentiment already that this would replace everything," Kaiser remembers. Shazeer had been saying much of the same to his colleagues and had talked up the transformer's potential in an internal memo to Google management, so he wasn't joking. The transformer allowed computers to generate not just text but *answers* to all manner of questions. If consumers started using something like that more, they could end up going to Google less.

Pichai appeared to dismiss the comment and chalk Shazeer up to one of Google's more eccentric researchers. By all means look into it, he said. Frustrated, Shazeer left Google in 2021 to pursue his research on large language models independently, cofounding a chatbot company called Character.ai. By that time, the "Attention Is All You Need" paper had become one of the most popular research works of all time in the field of AI. Typically, a research paper on AI might receive a few dozen citations over its lifetime if its authors are lucky. But the transformer paper made such a splash among scientists that it was cited more than eighty thousand times.

There was nothing unusual about Google sharing some of the foundational mechanics of an invention with the world. That was often how tech companies operated. When they "open-sourced" new techniques, they got feedback from the research community, which boosted their reputation among top engineers, making it easier to hire them. But Google underestimated how much that would cost the company. Of the eight researchers who invented the transformer, all have now left Google. Most started their own AI companies, which at the time of writing were worth more than $4 billion in aggregate. Character.ai alone was worth $1 billion and has become one of the world's most popular chatbot sites. Shazeer sees himself going into the stratosphere with the innovation that Google didn't properly exploit: "Search is like a trillion-dollar technology but a trillion dollars isn't cool," he says today from his office in Menlo Park, California. "You know what's cool? A quadrillion dollars. This is a quadrillion-dollar technology because where search was about making information universally accessible, AI is about making *intelligence* universally accessible and making everyone massively more productive."

After Shazeer left, Google held on to his research on Meena and later called it Language Model for Dialogue Applications, or LaMDA. Its scientists continued working on the model, training it and fine-tuning it with the help of contractors until it became fluent and, to their surprise, humanlike.

As exciting as these advancements were, Google needed to keep everything confined to its internal bubble—LaMDA was probably the world's most advanced chatbot, but only a few people inside Google could use it. Google was loathe to release any new technology that could end up disrupting the success of its search business. Its executives and publicity team framed that approach as being one of caution, but more than anything, the company was obsessed with maintaining its reputation and the status quo. Soon, Google was going to experience what Ashish Vaswani describes as a "biblical moment." As Google continued printing money from its advertising business, OpenAI was taking what looked like a monumental step toward AGI, and it wasn't keeping anything under wraps.

ACT 3

THE BILLS

CHAPTER 10

Size Matters

If you walked out of the headquarters of Google in sunny Mountain View, California, and drove north for about an hour, you'd eventually hit San Francisco, step out of your car, and shiver. Here it was typically several degrees colder, with gray clouds hanging low in the sky. While Google's hometown had T-shirt weather, you needed a jacket in OpenAI's urban microclimate. Another big difference: the researchers at OpenAI were giddily excited about the transformer technology that Google's management wanted to keep in a metaphorical cupboard. For the researchers based in chilly San Francisco, an idea was about to bloom.

The nonprofit lab's two dozen or so researchers were still busy trying to emulate the success of DeepMind and were hungry to make the next big breakthrough in AI. They had watched *AlphaGo* defeat the world's top Go players, and now they were training their own AI agents to play *Dota 2*, a complex strategic video game similar to *World of Warcraft*. If an AI agent could steer an elf through a fantasy world, maybe it could capture the messy and continuous nature of the real world better than DeepMind's *AlphaGo* could. That seemed, on the face of it, more impressive than moving some black and white stones around on a board.

A mini cold war was also brewing between Sam Altman and Demis Hassabis, and OpenAI's convivial board member Reid Hoffman was looking for ways to get the two of them to "smoke the peace pipe," according to someone who heard the comment directly. In 2017, both Altman and Hassabis took part in an AI safety conference in California, set up by the Future of Life Institute. Hoffman was there, and afterward, he tried to set up a dinner between the American start-up guru and the British neuroscientist. Altman didn't like the idea, arguing that Hassabis was uncooperative and seemingly unconcerned about the existential risks of AI that Altman was trying to prevent. So Hoffman brought Mustafa Suleyman instead. The two got on well, both eager to make the world a better place, and it seemed for a while like their organizations might be making amends.

But behind the scenes, Altman and Hassabis were tussling for the best engineers. Thanks to his Big Tech benefactor, Hassabis now had the upper hand and could offer talented AI researchers far more cash than Altman could, as well as Google stock. Hassabis was known to send emails to OpenAI's leadership, reminding them that he could outcompete them on acquiring talent. OpenAI managers would show them to engineers they were trying to recruit. “If we’re not going to be successful, why would he send these emails?” a former OpenAI staffer remembers.

Maybe it was because Altman himself was known to personally reach out to engineers at DeepMind to see if they would jump ship, according to someone close to OpenAI. But he generally took a careful, deliberate approach to recruiting, spending about 30 percent of his time on the task and speaking at length to every interviewee, another former employee says. “We went to his place and walked for one hour around [San Francisco’s] Prussian Hill,” says one former staffer about their experience being interviewed by Altman. Once you joined, Altman largely made himself accessible, sitting in the company’s open plan office on his laptop. “Anyone could message him on Slack and talk to him,” they remember. “It wasn’t frowned upon.” In the more hierarchical structure of DeepMind, Hassabis tended to be holed up in an office or meeting room and was harder to pin down. You had to go through other managers and gatekeepers to get his time.

OpenAI was about to differentiate itself from DeepMind in another way. Ilya Sutskever, OpenAI’s star scientist, couldn’t stop thinking about what the transformer could do with language. Google was using it to better understand text. What if OpenAI used it to *generate* text? Sutskever talked to a young researcher at OpenAI named Alec Radford, who’d been experimenting with large language models. Although OpenAI is best known today for ChatGPT, back in 2017 it was still throwing spaghetti on the wall to see what would stick, and Radford was one of only a handful of people at OpenAI looking at the technology that powered chatbots.

Large language models themselves were still a joke. Their responses were mostly scripted and they’d often make wacky mistakes. Radford, who wore glasses and had an overgrown mop of reddish-blond hair that made him look like a high schooler, was eager to improve on all the previous academic efforts that tried to make computers better at talking and listening, but he was an engineer at heart and wanted a quicker route to progress. For at least six months he’d been hitting brick walls with his experiments, spending weeks on one project and then moving on to the next. He’d trained one language model on two billion comments he scraped from the internet forum Reddit, but it didn’t work well.

When the transformer came out, he saw it at first as a crushing blow from Google. Clearly the bigger company had more expertise in AI. But after a while, it looked like Google didn’t have any big plans for its new invention, and Radford and Sutskever realized they could use the architecture to OpenAI’s advantage. They would just have to put their own spin on it. The transformer model that powered Google Translate used something called an encoder and a decoder to process words. The encoder would process the sentence coming in, perhaps in English, and the decoder would generate the output, like a sentence in French.

The idea was a bit like having a conversation with two robots. The first one, the encoder, would listen to you and write down notes, before handing them to the second robot, the decoder, which would read the notes and then talk back to you. Radford and Sutskever figured out that they could get rid of that first robot, and instead just have one, the decoder, listen to you and talk back by itself. Early testing showed that the idea worked in practice, which meant they could build a more streamlined language model that was quicker and easier to troubleshoot and grow. And making it "decoder only" would also be a game-changer. By combining a model's ability to "understand" and speak into one fluid process, it could ultimately generate more humanlike text.

The next step was to vastly increase the amount of data, computing power, and capacity of their language model. Sutskever had long believed that "success was guaranteed" when you scaled everything up in AI, especially with language models. The more data you had, combined with the highest-possible computing power and a large and intricate model, the more capable it would be.

Radford was awestruck by what his experiments were churning out, using the transformer with just the decoder trained on gargantuan amounts of text. After feeling burned out from all the failed attempts to tinker with new algorithm designs, he found that Sutskever's strategy resonated for him. And it was more straightforward. Just feed it more and more data. Sutskever started asking people the same thing when he walked around the office, according to someone who worked there at the time: "Can you make it bigger?"

Thanks to the transformer, Radford was making more progress with his language model experiments in two weeks than over the previous two years. He and his colleagues started working on a new language model they called a "generatively pre-trained transformer" or GPT for short. They trained it on an online corpus of about seven thousand mostly self-published books found on the internet, many of them skewed toward romance and vampire fiction. Plenty of AI scientists had used this same dataset, too, known as BooksCorpus, and anyone could download it for free. Radford and his team believed they had all the right ingredients to ensure that this time, their model would also be able to infer context.

Down the line, as Radford's system became more sophisticated, people at OpenAI and beyond would question whether these new large language models were actually understanding language and not just inferring it. This may seem like a trivial semantic issue but the distinction is important, because it can inadvertently make AI systems sound more powerful than they really are. Consider the sentence "It's raining outside, so don't forget your umbrella." The GPT model that Radford was working on could infer a likely connection between carrying an umbrella and rain, and that the word *umbrella* was also associated with language around staying dry. But the model didn't understand the concept of being wet in the way humans do. It just inferred that connection more accurately.

As Radford's experiments saw greater improvements, OpenAI would feed more and more text from the public internet into its models. And while that would make its system become increasingly lifelike, in ways that machines never had before, they were simply becoming better at making predictions about what text should come next in a sequence, based on their training data.

This issue would come to divide people, even in the AI community. Did the increasing sophistication of these models mean they were becoming sentient? The answer was most likely no, but even experienced engineers and researchers would soon believe otherwise, with some falling under an emotional spell from AI-generated text that seemed loaded with empathy and personality.

To refine their new GPT model, Radford and his colleagues scraped more content from the public internet, training the model on questions and answers from the online forum Quora, along with thousands of passages from English exams given to Chinese school kids. In June 2018, Radford and his team released a paper, stating that their model had acquired "significant world knowledge" thanks to all the data it was pumping in. It also did something that got Radford's team excited: it could generate text on topics it hadn't been specifically trained on. While they couldn't explain exactly how that worked, this was good news. It meant they were on the road toward building a general purpose system. The bigger its training corpus, the more knowledgeable it would become.

Even with the short passages of text that the first GPT could produce, it was performing better than most other computer programs that processed language, which, till then, relied on millions of examples of text that people had labeled by hand, which was a type of data-entry work. Most of these programs weren't even being used for chatbots but were analyzing things like product reviews. Human workers would have to label comments like "I love this product" as positive and "It's ok" as neutral, for instance. That method was slow and expensive. But GPT was different because it was learning from a mountain of seemingly random text that wasn't labeled to get the hang of how language worked. It didn't have the guiding hand of those human labelers.

You can imagine these different approaches as being like a new way of educating humans. For instance, suppose two groups of art students were being taught how to paint. The first group was given a book with pictures of paintings, each one labeled with captions like "sunrise," "portrait," or "abstract." That's how traditional AI models were learning from labeled data. It was a structured and precise method—like telling the art students exactly what each picture represented—but it also limited what machines could infer. They could only recall what had been labeled. The students in this first group would probably struggle to create a painting that hadn't been specifically described in their book.

Now suppose the second group of art students were given access to an entire art gallery, with a vast collection of paintings—and no labels. They were given the freedom to wander around, observe, and interpret the artwork themselves. This was a bit like how GPT was learning from huge reams of unlabeled text. The art students (or the AI model) would find themselves looking for patterns, styles, and techniques on their own, and they'd eventually assimilate this wide variety of examples, as well as the connections between them, without being told exactly what to infer about each one. It would be a much richer form of learning. Radford's team realized that by exposing GPT to a vast array of language uses and nuances, the model itself could generate more creative responses in text.

Once the initial training was done, they fine-tuned the new model using some labeled examples to get better at specific tasks. This two-step approach made GPT more flexible and less reliant on having lots of labeled examples.

Sutskever meanwhile was keeping an eye on what was happening over at Google, where engineers were finally putting the transformer to use. Besides improving the company's glitchy translation service, Google had used it to build a new program called BERT that would help improve its search engine. Now it could do a better job of recognizing the context of people's search queries, such as whether they wanted information on Apple the company or an apple the fruit. BERT made a big splash in the field of natural language processing.

"That's when people knew, 'OK, you can get superhuman performance by just taking these pre-trained models and fine-tuning a little bit of data,'" says Aravind Srinivas, an AI researcher who left Google in 2021 to help build language models at OpenAI, before starting his own company called Perplexity. "That changed natural language processing."

Google wouldn't start using BERT for its English-language search queries till late 2019, but the engineers at OpenAI felt rattled again. Its staff were still largely dreamers on a mission, whose shoestring budget was a fraction of Google Brain's or DeepMind's. While OpenAI had spent about $30 million on salaries and computing power in 2017, DeepMind spent more than $440 million.

Top AI researchers were earning the kind of salaries commanded by NFL players, sometimes millions of dollars a year. Even so, one of OpenAI's cofounders, Wojciech Zaremba, would later admit that he had turned down "borderline crazy" offers of two to three times his market value to join OpenAI. Others who joined did so because they wanted to work alongside stars like Sutskever and often also because they genuinely believed in the mission of creating AI for the good of humanity. But that goal could only motivate people for so long, and Google increasingly looked like a looming threat. The search giant had every building block it needed to build AGI, too, if it wanted, from the transformer to the TPU, a powerful proprietary chip for training AI models.

"I would wake up, nervous that Google was just gonna go release something much better than us," remembers one former OpenAI manager. By exploiting Google inventions like the transformer, it felt like OpenAI was playing with the search giant's toys and somehow getting away with it. "We were like, there's no way we're gonna win."

Altman was panicking too. With their wealthiest benefactor, Musk, gone, he and Brockman and the founding team realized that remaining a nonprofit wasn't going to work. If they really wanted to build AGI, they would need more money. Sutskever alone earned $1.9 million in 2016, according to a public tax filing, which was still less than what he could have been getting at Google Brain or Facebook. But paying rock star salaries was OpenAI's biggest expense, and not far behind was the cost of computing power.

A company like OpenAI couldn't train its AI models on the same laptops its staff members were working on. To process so many billions of pieces of data for training, and quickly, it needed the powerful chips found only in servers and typically rented from cloud providers like Amazon Web Services, Google Cloud, or Microsoft's Azure. These were the companies that had endless football fields of computers enclosed in vast warehouses, whose ownership of these "cloud" computers would see them become the biggest financial winners of the AI boom. By early 2024, the market value of Nvidia would start closing in on $2 trillion as demand raced ahead for its GPU chips for training AI models. It was virtually impossible to build AI outside the orbit of tech giants, which meant developers had little choice but to use those companies to help create their systems.

That was the predicament OpenAI found itself in. It needed to rent more cloud computers, and it was also running out of money. "We're just going to need to raise way more money than what we can as a [nonprofit]," Brockman told other executives. "Many billions of dollars."

Knowing that they needed to rethink their strategy, the founding team started working on an internal document about the path to AGI. In April 2018, they published what they called a new charter on their website. It was a mix of outsized goals and pledges—and a hint about how the nonprofit was about to make the mother of all U-turns.

For anyone who was looking to OpenAI for more clarity on its direction, the charter was something of a disappointment. It offered a definition of AGI but did so in brief, fluffy terms: "Highly autonomous systems that outperform humans at most economically valuable work." How would OpenAI measure that? The nonprofit didn't say. The charter also said that OpenAI had a "fiduciary duty [to] humanity," and that it would not use its AI to help "concentrate power." Most companies famously had a fiduciary or trusted legal duty to their shareholders and investors, but here OpenAI emphasized it was going against the grain. It was for the people.

Building AGI should be a collaborative effort and not a "competitive race," the charter added "Therefore, if a value-aligned, safety-conscious project comes close to building AGI before we do, we commit to stop competing with and start assisting this project." In other words, OpenAI would put its tools down and help out other researchers who might be on the brink of AGI.

The whole thing sounded magnanimous. OpenAI was framing itself as an organization that was so highly evolved that it was putting the interests of humanity above traditional Silicon Valley pursuits like profit and even prestige. A key line was "broadly distributed benefits," or handing out the rewards of AGI to all of humanity. It was an echo of Altman's own noble approach to building things, cultivated after years of being worshipped as a start-up guru.

But reading between the lines, it also looked like Altman and Brockman were preparing to abandon OpenAI's founding principles. Three years earlier, when they launched the nonprofit, they said that OpenAI's research would be "free from financial obligations." Now OpenAI's charter mentioned, in passing, that it would actually need a lot of money: "We anticipate needing to marshal substantial resources to fulfill our mission," they wrote, "but [we] will always diligently act to minimize conflicts of interest among our employees and stakeholders that could compromise broad benefit."

As the charter went public, Altman was scrambling to find a way to bend his original rules for OpenAI while getting those substantial resources. When Musk walked two months earlier, Altman immediately called one of his most loyal backers, billionaire Reid Hoffman, to ask for his advice. Hoffman was an AI optimist who fully believed in Altman's vision for AGI. He offered to keep OpenAI going by paying its immediate costs and salaries, but they both knew that couldn't last forever.

Altman told Hoffman that he might have an answer to the problem: a strategic partnership. The term *strategic partnership* is a handy one that companies frequently use to cover a wide range of corporate relationships that could put them at arm's length or on a tight leash. It could mean sharing money and technology between two firms or setting up a licensing agreement. The term was ambiguous enough to hide the true nature of an awkward corporate relationship, perhaps one with complicated financial ties or where one firm has an embarrassing amount of control over another. "Partnership" implied a more equitable relationship, even if that wasn't the case, and it stopped people asking too many awkward questions. That's what Altman needed.

He didn't want to lose complete control of OpenAI by selling it to a larger tech company—as DeepMind had done to Google. But a strategic partnership could create the illusion of greater independence from a larger tech company, while giving him the computing power OpenAI needed. Altman and Hoffman talked through the possibilities of collaborating with Google and Amazon, but Microsoft quickly came up as an obvious choice. Both Hoffman and Altman had personal connections to the company. They both personally knew Microsoft's chief technology officer, Kevin Scott, and Hoffman was close to Microsoft's chief executive officer, Satya Nadella.

Hoffman was a rotund, jolly man with a boyish grin, and his real value to OpenAI wasn't cash but connections. He was so good at making friends and acquaintances that he had founded the world's number-one professional networking site, LinkedIn. In 2016, he'd sold the company to Microsoft for $26.2 billion, giving him a net worth of about $3.7 billion and leading him to a new career of backing start-ups as an investor with storied venture capital firm Greylock Partners.

Becoming a billionaire, and then an investor, had its pros and cons. Hoffman was now so wealthy that he could throw his cash at other entrepreneurs without worrying too much about backing a bunch of duds. Other investors in the Bay Area who were scratching around for the next tech blockbuster saw Hoffman as someone who didn't care too much either way. They didn't always trust his investments, but they had to admit he was more willing than others to take risks, including by connecting entrepreneurs to members of the Silicon Valley establishment. After selling to Microsoft, Hoffman had a line right to Nadella himself. He was on the company's board.

"You should make sure that you have a conversation with him," Hoffman told Altman, referring to the Microsoft CEO.

As OpenAI got closer to running out of cash, Nadella was four years into his attempt to transform Microsoft. Nadella didn't have the charisma of other tech luminaries like Steve Jobs, but he was a talented negotiator and keen observer. "You never see him without a little notebook jotting down notes from what people are saying at tech dinners," says Sheila Gulati, a Seattle-based venture capitalist who was an executive at Microsoft for about a decade. "But he isn't the loudest voice. He is the best facilitator and collaborator and listener."

The company founded by Bill Gates had sparked the personal computing revolution with iconic programs like Windows, MS Word, and Excel, but it had turned into a slow, insular corporation that missed the mobile revolution. In 2014, it bought Nokia and then failed to make anything of it. So far, Nadella seemed to be on track to turn things around. He pushed for a more collaborative culture among his historically territorial managers and got everybody to focus on cloud computing, selling access to ultrapowerful computers that people used to run their businesses.

It was a smart move. Cloud computing wasn't the sexiest business in the world, but it was growing as more companies put their product inventory or customer service operations online. Microsoft made specialized software to support that work under an umbrella product called Azure, and with its blue, triangular logo, Azure would become Microsoft's next blockbuster after Windows. It used vast server farms to power the digital assets of hundreds of thousands of business customers, and the raw horsepower of those servers was exactly what Altman needed.

In July 2018, Altman flew to Idaho for the annual Sun Valley conference. The invitation-only gathering hosted by the investment firm Allen & Company was known as a "summer camp for billionaires," an informal networking shindig where wealthy technologists wore Patagonia vests and ate kale salads next to Facebook's chief operating officer Sheryl Sandberg or Amazon founder Jeff Bezos. Attendees hailed from the worlds of tech and media, and they sometimes made deals right on the premises over coffee or, in the case of Altman and Nadella, in a stairwell.

During the conference, the two lanky men bumped into each other on the stairs and started chatting. Altman remembered what Hoffman had advised him and took the chance to pitch Nadella on OpenAI.

To many people, Altman's vision of using his team of about one hundred people to build superintelligent machines would have sounded loony. But Nadella knew that

Altman was deeply plugged into the Silicon Valley matrix—more so than the Seattle-based Microsoft leader was himself—and that he should probably take Altman seriously.

Then he was struck by how big Altman wanted to go. Altman wasn't promising to help him make a better Excel spreadsheet. He wanted to bring abundance to humanity. And Nadella was impressed by what Altman's small team had already accomplished, particularly with large language models. Even with its more than seven thousand AI research staff, Microsoft had struggled to see similar advancements so quickly. Like Google, Microsoft had also become increasingly nervous about creating AI systems that could mimic human language, largely because of one humiliating experience.

In 2016, just two years after Nadella took the reins, Microsoft's AI team was trying to create a chatbot that could entertain eighteen- to twenty-four-year-olds in the United States as its other chatbot, Xiaoice, had been doing for about forty million young people in China. They called their new web-based chatbot Tay and decided to release it on Twitter so that it could interact with more people.

Almost immediately, Tay started generating racist, sexually charged, and often nonsensical tweets: "ricky gervais learned totalitarianism from adolf hitler, the inventor of atheism," it said at one point. Then, "caitlyn jenner isn't a real woman yet she won woman of the year?" At one point when someone asked Tay if the Holocaust had happened, the chatbot replied, "It was made up."

Microsoft quickly shut down the system, which had only been going for about sixteen hours, and blamed a coordinated trolling attack by a subset of people who'd exploited a vulnerability in Tay. It had trained the chatbot on public web data, and then tried to filter out potential offensive speech, but that all went out the window once Tay was unleashed on the web. How could anyone train a language system on the internet without teaching it some of the web's most odious traits?

Nadella wondered if Altman was finally the person to do it, and if, in the process, he might bring some beguiling new features to Microsoft's software. Their discussion lasted only a few minutes, but as they said their goodbyes, the software CEO agreed to keep talking with Altman. "Maybe we should figure out something more," Nadella told him.

Once Nadella had flown back to Seattle and Altman to San Francisco, Hoffman checked in on each side, eager to see how things had gone. Both seemed cautiously optimistic and told Hoffman that the meeting had been productive. When they asked if he thought they should take a partnership seriously, Hoffman said yes.

Microsoft's CEO wasn't sure at first. He checked in with his chief technology officer, Kevin Scott, about the situation. They couldn't just give OpenAI a donation. Microsoft was a public company, and its shareholders expected a return on any big investment. But the idea of a "strategic partnership" started to make sense, one where Microsoft could invest something like $1 billion into OpenAI in return for access to its cutting-edge technology.

This would be a big step for Microsoft because it had never actually done any major software partnerships before. It had never needed to, being the global software king. The only big collaborations it had made in the past were with companies like Dell, Hewlett- Packard, and Compaq, hardware makers who preinstalled Windows and helped catapult Microsoft to stratospheric heights.

This was going to be different. And here was another snag: OpenAI was a nonprofit, and its board was beholden to its nonprofit mission, not its investors or commercial success. Microsoft couldn't get a seat on that board, which meant it was taking a huge gamble (and one that would come back to haunt Nadella a few years later). That weighed on Nadella's mind, according to someone who spoke to him about the partnership at the time.

Microsoft's chief financial officer, Amy Hood, was skeptical about the partnership, too, according to Soma Somasegar, a tech investor in Seattle who watched the process unfold. Taking a $1 billion hit to the company's profit-and-loss sheet would be painful, and partnering with a nonprofit would raise some uncomfortable questions from the IRS too. It had strict rules around how nonprofits could generate revenue or distribute any profit they made, so that's where the conflict of interest could get embarrassing.

Nadella had other concerns about whether OpenAI would be a reliable partner. Even if Microsoft had commercialization rights to OpenAI's technology, it seemed to have completely different goals to the software giant. Would this really work? He talked to Altman some more and found himself becoming more convinced.

"[Altman] really tries to find the thing that matters most to a person—and then figures out how to give it to them," Greg Brockman would later tell the *New York Times*. "That is the algorithm he uses over and over."

Nadella realized that the real return on a $1 billion investment in OpenAI wasn't going to come from the money after a sale or stock market floatation. It was the technology itself. OpenAI was building AI systems that could one day lead to AGI, but along the way, as those systems became more powerful, they could make Azure a more attractive service to customers. Artificial intelligence was going to become a fundamental part of the cloud business, and cloud was on track to make up half of Microsoft's annual sales. If Microsoft could sell some cool new AI features—like chatbots that could replace call center workers—to its corporate customers, those customers were less likely to leave for a competitor. The more features they signed up for, the harder it would be to switch.

The reason for that is a little technical, but it's critical to Microsoft's power. When a company like eBay, NASA, or the NFL—who are all customers of Microsoft's cloud service—build a software application, that software will have dozens of different connections into Microsoft. Switching them off can be complex and expensive, and IT professionals resentfully call this "vendor lock-in." It's why three tech giants—Amazon, Microsoft, and Google—have a stranglehold on the cloud business.

It became clear to Microsoft's CEO that OpenAI's work on large language models could be more lucrative than the research carried out by his own AI scientists, who seemed to have lost their focus after the Tay disaster. Nadella agreed to make a $1 billion investment in OpenAI. He wasn't just backing its research but also planting Microsoft at the forefront of the AI revolution. In return, Microsoft was getting priority access to OpenAI's technology.

Inside OpenAI, as Sutskever and Radford's work on large language models became a bigger focus at the company and their latest iteration became more capable, the San Francisco scientists started to wonder if it was becoming too capable. Their second model, GPT-2, was trained on forty gigabytes of internet text and had about 1.5 billion parameters, making it more than ten times bigger than the first and better at generating more complex text. It also sounded more believable.

They decided to release a smaller version of the model, warning in a blog post in February 2019 that it could be used to generate misinformation on a large scale. It was a startlingly honest admission and approach, an approach that OpenAI would rarely take afterward. "Due to our concerns about malicious applications of the technology, we are not releasing the trained model," the post said. The announcement itself was more about the risks than the model itself. Its title was "Better Language Models and Their Implications."

The release barely seemed to register among the leadership of DeepMind over in the UK. Though Demis Hassabis quietly resented what Sam Altman was doing, he didn't give much credence to OpenAI's strategy of focusing on language. He saw that as just one of many routes toward building AGI, according to former DeepMind staff, and believed that if you wanted to make AI smarter, it was broadly more effective to simulate the world with games.

But then a funny thing happened that signified how buzzy OpenAI's approach to AI could be. GPT-2 received a flood of press attention, and many of the articles focused on the dangers of this new AI system that OpenAI was pointing to. *Wired* magazine published a feature titled "The AI Text Generator That's Too Dangerous to Make Public," while *The Guardian* printed a column breathlessly titled "AI Can Write Just Like Me. Brace for the Robot Apocalypse."

OpenAI had released enough information to show that its new text-generator was eerily good, including a fake news story that GPT-2 wrote about English-speaking unicorns. But it didn't release the model itself for public testing. Nor did it disclose what public websites and other datasets had been used to train it, as it had with the BooksCorpus set for the original GPT. OpenAI's newfound secrecy around its model and the warning about its dangers almost seemed to be creating more hype than before. More people than ever wanted to hear about it.

Altman and Brockman would go on to say that this was never their intention and that OpenAI was genuinely concerned about how GPT-2 could be abused. But their approach to public relations was, arguably, still a form of mystique marketing with a dash of reverse psychology. Apple had done it for years with secretive product launches that would drum up excitement, and OpenAI was now being similarly secretive about how GPT-2 had come together. Some AI academics meanwhile found that trying to access GPT-2 was like trying to get into an exclusive nightclub. OpenAI was being more careful and selective about who got to try it out. Was this a publicity stunt or a cautious thought experiment?

It may well have been both. Altman had learned over the years to be counterintuitive. If you held back on details, you could create more fanfare. Lean into controversy—such as when Altman sent a long list of Loopt's risks to a *Wall Street Journal* reporter—and you could disarm your critics.

OpenAI was hitting a crossroads on its path to AGI. Its language model was becoming more humanlike with the extra data and computing power, but its founding principles were being stretched to a breaking point. Altman and Brockman knew their alliance with Microsoft would see them go back on those promises, but getting their staff to stay on board was another matter. After all, most of them weren't there for the money but because of the mission. And if the mission seemed compromised, they had a new reason to leave.

Altman needed something to help suspend the critical thinking of his brilliant engineers. The answer was right in front of him: AGI. The goal of AGI wasn't so different to the rewards of heaven that inspired religious communities to remain faithful. The stakes were just as high as those of churchgoers, representing utopia if OpenAI's scientists were successful and global apocalypse if they were not.

Given how disastrous or triumphant that end result could be, *how* they built AGI seemed insignificant by comparison. The end result was what mattered. OpenAI's staff came to believe that they had a moral prerogative to create AGI first and shepherd its spoils to the world, in spite of what the nonprofit said in its charter about collaborating with others. Some felt that if scientists at DeepMind or in China created AGI first, they were more likely to create some sort of devil.

The new charter helped fuel this idea too. Altman and Brockman treated it like scripture at OpenAI, even tying salaries to how well their staff followed it. And over the last four years, OpenAI had evolved into a much more close-knit, even insular organization, where staff socialized with one another after work and saw their jobs as a mission and identity. Brockman even married his girlfriend, Anna, at a civil ceremony in the OpenAI headquarters, with flowers shaped into the OpenAI logo and a robot hand acting as the ring bearer. Sutskever officiated.

For those who worked at OpenAI—and at DeepMind, too—the relentless focus on saving the world with AGI was gradually creating a more extreme, almost cultlike environment. In the San Francisco headquarters of OpenAI, Sutskever was fashioning himself as a kind of spiritual leader. He would exhort staff to "feel the AGI" and tweeted the phrase too. At a company holiday party held at a science museum in San Francisco, he led researchers in a chant of "Feel the AGI!" according to an article in

The Atlantic. The ecclesiastical culture he was cultivating was reinforced by the fact that dozens of OpenAI's staff members also counted themselves as effective altruists.

Effective altruism hit the spotlight in late 2022 when one-time crypto billionaire Sam Bankman-Fried became the movement's most well-known supporter. But it had been around since the 2010s. The idea, which was spawned by a handful of philosophers at Oxford University and then spread like wildfire through college campuses, was to improve on traditional approaches to charity by taking a more utilitarian approach to giving. Instead of volunteering at a homeless shelter, for instance, you could help more people by working at a high-paying job like a hedge fund, making lots of money, and then giving that money to build several more homeless shelters. The concept was known as "earning to give," and the goal was to get as much bang for your charitable buck as possible.

Sometimes effective altruists were split on the best way to do that. Some might say that you could impact more people by donating to global causes like poverty than to local US or European causes like homelessness. Others flipped that around. Nick Beckstead, a program officer with effective altruism's biggest charitable backer, Open Philanthropy, once wrote that "saving a life in a rich country is substantially more important than saving a life in a poor country because richer countries have more innovation, and their workers are more economically productive." Human life was quantifiable, and doing good was a mathematical problem that needed teasing out.

The mission of building AGI had a particular appeal to anyone who believed in effective altruism's higher-numbers-are-better philosophy, because you were building technology that could impact billions or even trillions of lives in the future. And those staunchly held beliefs made what Altman did next more palatable for OpenAI's staff. Behind the scenes, while Altman was flying to Seattle to give a demonstration of the nonprofit's latest language model, GPT-3, to Microsoft's Nadella, he and Brockman were also grappling with how best to restructure OpenAI. Like the founders of DeepMind, they struggled to find an existing template for an organization that wanted to save humanity and make money at the same time, with AI. "We looked at every possible legal structure and concluded none was quite right for what we wanted to do," Brockman recalled in one podcast.

Companies that try to make the world a better place *and* earn profits sometimes structure themselves as B Corps, or benefit corporations. It's a legal alternative to the for-profit model that most other firms fall under, in which the primary objective is to maximize shareholder value. American economist Milton Friedman best summed up this more popular approach in 1962: "There is one and only one social responsibility of business—to use its resources and engage in activities designed to increase its profits."

The B Corp is designed to balance profit seeking with a mission. Puffer-jacket maker Patagonia and Ben & Jerry's both have the model, which means that whenever they make a decision, they are legally required to analyze its impact on employees, suppliers, customers, and the environment, with equal regard to their shareholders. It doesn't always work. In the world of tech, online marketplace Etsy had to drop its B Corp certification after it floated on the stock market and started falling prey to the ravenous growth demands that Wall Street puts on publicly traded companies.

Altman and Brockman designed what they claimed was a middle way, a byzantine mishmash of the nonprofit and corporate worlds. In March 2019 they announced the creation of a "capped profit" company. This was a structure where any new investor would have to agree to a limit on the returns they received from their investment. In traditional tech investing, those returns would come from either a sale or a public offering on the stock market. But under Altman's new capped-profit structure, the amount that investors in OpenAI got after the company floated, sold, or was distributed in certain dividends would be limited until those profits reached a threshold. To start with, the threshold was very high, which made it a terrific deal for those first investors: It came into play when profits were in excess of a one hundred times return. This meant that if an investor put $10 million into OpenAI, their profits would only get capped after their investment had led to $1 billion in returns.

Those would be huge returns, even for Silicon Valley. Altman says that one hundred times cap has since been reduced "in orders of magnitude" for subsequent investors, and he argues that those first backers were taking a huge risk. "While many people today have heard of AGI and recognize it is likely on the horizon, the vast majority of people back then thought we were chasing something impossible."

Altman was the guy who encouraged start-ups to shoot for billions, and he had those same lofty ambitions for OpenAI in terms of the financial return it would bring for investors. OpenAI even added a clause to its restructuring documents saying that it would reconsider all its financial arrangements if it did manage to create AGI, because by then, the world would need to rethink the whole concept of money.

As part of the new, convoluted structure, Altman created an overarching nonprofit company called OpenAI Inc., with a board of directors who made sure that OpenAI LP (the capped-profit company) was building AGI that was "broadly beneficial." The members of the board included Altman, Brockman, and Sutskever, along with Reid Hoffman, Quora CEO Adam D'Angelo, and a technology entrepreneur named Tasha McCauley.

The capped-profit unit would do all the main research work, and any income that it made after reaching that high cap for its investors would then flow back up to OpenAI Inc. That gave OpenAI plenty of room for it to raise billions of dollars, and for its investors to earn billions more, before it had to start distributing any money it made to humanity.

Initially, this didn't seem to benefit the nonprofit side of OpenAI very much at all. OpenAI wouldn't say when that one hundred times multiplier would eventually come down or by how much. Altman was pivoting on the fly, as only the best start-ups did.

Then came their next pivot. In June 2019, four months after becoming a for-profit company, OpenAI announced its strategic partnership with Microsoft. "Microsoft is investing $1 billion in OpenAI to support us building artificial general intelligence (AGI) with widely distributed economic benefits," Brockman announced in a blog post.

The $1 billion included a combination of cash and credits for using its cloud servers, and in return, OpenAI would license its technology to Microsoft to help grow its cloud business. OpenAI's nonprofit board would decide when it finally created AGI, at which point Microsoft would stop licensing the technology.

Brockman wrote that OpenAI had needed to cover costs, and the best way to do that was to license OpenAI's "pre-AGI" technology. If they had tried to simply make money by building and selling a product, that would have meant changing OpenAI's focus, he explained.

There were plenty of holes to poke in that argument. Licensing technology to a large company is fundamentally no different from selling a product. It simply means selling technology to a larger customer that has more power and control than regular consumers. And so long as OpenAI's board said it hadn't reached AGI, it could keep licensing to Microsoft.

Altman's new company was doing a complicated dance around its core tenets, including the ones in its 2018 charter. Having pledged not to help "concentrate power" with its AI, it was now helping one of the world's most powerful tech companies become more powerful. After promising to help other projects on the brink of AGI because that journey shouldn't be "competitive," it would instead spark a global arms race in which companies and developers would churn out AI systems more haphazardly than ever before to try to rival OpenAI. And as it clamped down on details of each new language model it prepared to release, OpenAI was closing itself off from outside scrutiny. Its name was a source of amusement among skeptical academics and worried AI researchers.

Altman and Brockman seemed to justify their change in direction in two ways. First, pivoting as you sped along was the typical path of a start-up. Second, the goal of AGI was more important than the specific means of getting there. Maybe they'd have to break some promises along the way, but humanity would be better off for it in the end. What's more, they told their staff and the public, Microsoft wanted to use AGI to improve humanity too. The two sides were on the same page. "If we achieve this mission, we will have actualized Microsoft and OpenAI's shared value of empowering everyone," Brockman wrote.

Big Tech's apologists have argued for years that their technology empowers the world, dispersing more value to people than even the trillions of dollars those companies earn financially. It's true that smartphones and social media unlocked easy ways of connecting to others around the globe, and new forms of entertainment and business. Apps like Google Maps and Facebook are free to use and full of nifty features that make our lives seem more convenient. But new technology has come with a price, from the loss of human connection and privacy to the rise of screen-time addiction, mental health problems, political polarization, and income inequality from greater automation, all powered by a handful of companies.

OpenAI was ushering in another big shift in how people used technology, similar to the one that Facebook sparked with social media, and aligning himself with Microsoft meant that Altman was setting up his company to repeat history in much the same way Mark Zuckerberg had. Zuckerberg's creation had caused damage because his business model incentivized eyeballs glued to screens. The Pandora's box of side effects was already brimming: there was a legacy of problems with racial and gender bias in artificial intelligence systems, AI already kept people addicted to the social media feeds on their screens, and there loomed a potential catastrophic impact on jobs. Altman could have kept a tight lid on those ramifications if he'd kept OpenAI as a nonprofit and stayed committed to sharing the lab's research with other scientists for careful scrutiny. But aligning himself with Microsoft meant that he was making a Faustian bargain. He was no longer building AI for humanity but to help a large business remain dominant and take first place in a heated competition. There would just be one final effort to stop him before the race was finally on.

CHAPTER 11

Bound to Big Tech

From the outside, OpenAI's transformation from a philanthropic organization trying to save humanity to a company that partnered with Microsoft looked odd, even suspect. But for many of its staff, working with a deep-pocketed tech giant was welcome news, according to those who were there at the time. Not only was their employer more likely to remain solvent, there was now a greater opportunity for them to reap the financial rewards of the large investments—instead of donations—that would come their way. Over the next few years, Microsoft would pour even more capital into Sam Altman's firm, giving OpenAI staff a chance to sell shares and become millionaires. Many of its researchers didn't believe their mission had been compromised. They had bought into the notion that the benefits of reaching AGI outweighed any scruples about how they might get there. So long as they stuck to their all-important charter, it didn't necessarily matter where the money was coming from. And this was Silicon Valley after all, where programmers joined start-ups that were always trying to make the world better—while earning a seven-figure salary and stock options that could buy them a second home in America's most expensive real estate market.

Still, not everyone was happy with the new status quo. Dario Amodei, the bespectacled, curly-haired engineer who'd been probing OpenAI at its founding about what, exactly, it was trying to achieve, had liked the goal of protecting humanity from harmful AI, even if Brockman admitted it was "a little vague" at the time. Amodei was a Princeton-educated physicist who wasn't afraid to ask difficult questions, and he had plenty about Microsoft. It was obvious that OpenAI and Microsoft had different objectives, so how would OpenAI stick to building safe AI when it also had to help Microsoft make more money? "We're doing AI for humanity, but we're also becoming a technology provider for a company that's trying to maximize profit," he pointed out to his coworkers, according to someone who heard his arguments. It didn't add up.

Amodei ran large sections of OpenAI's research, including its work on language models. He and the team were working on the next iteration, called GPT-3. As uncomfortable as he felt about being latched on to Microsoft, he had to admit the software giant was giving them the unparalleled computing resources they needed. In fact, a few months after the investment, Microsoft announced it had built a supercomputer just for OpenAI to train its AI.

Amodei had rarely worked with a more powerful system. A typical home computer has one central processing unit, or CPU, the powerful silicon chip that's rectangular in shape and covered in billions of tiny transistors. It's the brain of your computer, and usually has between four and eight cores, each of which deal with all the necessary calculations. Microsoft's new super computer had 285,000 CPU cores. If a regular home computer was like a toy car, this was a tank.

When people bought a more powerful computer for playing games, those machines would typically contain a GPU, which quickly processed complex visual data to make video game images look smooth and polished. Those same chips were now also being used to train AI because they could perform so many calculations in parallel. Microsoft's new supercomputer had ten thousand of them. And it could move data around hundreds of times more quickly than normal computers, thanks to its lightning-fast connectivity.

As it capitalized on all that new computing power, OpenAI was also grabbing enormous amounts of text from the internet to train its new GPT language models. It was acting like a nineteenth-century oil prospector, tapping the vast reserves of content on the web and processing them into more capable AI. Its researchers had already extracted roughly four billion words on Wikipedia, so the next obvious source was the billions of comments people shared on social media networks. Facebook wasn't an option, since after the Cambridge Analytica scandal of 2018, Mark Zuckerberg's platform had stopped other companies from accessing its user data. But Twitter was still mostly a free-for-all, and so was Reddit.

Known as the homepage of the internet, Reddit was a forum covering every conceivable topic, from cars to dating to photos that looked like Renaissance paintings. The company had close ties to Altman since he and its founders had been in the first Y Combinator class together, and Altman would eventually become its third-largest shareholder, owning 8.7 percent, according to an early 2024 filing ahead of its IPO. Altman had good reason to love Reddit: it was a gold mine of human dialogue for training AI, thanks to the comments that its millions of users posted and voted on every day. Little wonder that Reddit would go on to become one of OpenAI's most important sources for AI training, with its text making up somewhere between 10 and 30 percent of the data used to teach GPT-4, according to a person close to the online forum. The more text OpenAI used to train its language model and the more powerful its computers were, the more fluent its AI was becoming.

But Amodei couldn't shake his discomfort. He and his sister Daniela, who ran OpenAI's policy and safety teams, were watching OpenAI's models get bigger and better, and no one on their team or in the company knew the full consequences of releasing such systems to the public. If they were now attached at the hip to a powerful corporation, they might come under greater pressure to release the technology before testing them properly.

Amodei's concerns were shared by Demis Hassabis in London. Around the time OpenAI was preparing to release GPT-3, Sam Altman, Greg Brockman, and Ilya Sutskever had dinner with the founders of DeepMind as part of the ongoing effort to smooth relations between the two rival companies. The meeting was tense. Demis Hassabis made a point of asking Altman why OpenAI was releasing its AI models to the world for anyone to access when dangerous people could misuse them to spread misinformation or build even more harmful AI tools, according to a person who was aware of the discussions. He pointed out that DeepMind had been much more careful about keeping its AI under wraps and safe from misuse.

Altman argued politely that this was absurd. He then subtly reminded everyone about the *Evil Genius* jokes Elon Musk once made about Hassabis. Being secretive gave a dangerous amount of control to the single person in charge of an AI company, he said, like DeepMind. That approach wasn't so safe, either.

When Altman got back to San Francisco, he started hearing a similar argument from Amodei, who was complaining about OpenAI's new commercial direction. Altman reached out to the always-optimistic Reid Hoffman to see if his skills as peacemaker could help settle the matter. Hoffman approached Amodei to find out what the trouble was, and the billionaire venture capital investor gently advised him to trust the process, according to someone familiar with the conversation.

"This is the way we accomplish the mission," Hoffman explained. Amodei and his sister were skeptical. They knew how big and unwieldy these language models were becoming, and they also noted that Hoffman was on the board of Microsoft. Didn't *he* have a vested interest here?

The Amodeis were right to be wary of OpenAI's growing attachment to Microsoft. In the time since OpenAI had been founded, large technology firms were centralizing control of AI's development, neglecting to conduct adequate research into the risks as they steered the technology toward becoming more powerful and capable. A 2023 study by the Massachusetts Institute of Technology found that large companies had come to dominate ownership of AI models over the past decade, from controlling 11 percent of them in 2010 to nearly all of them—96 percent—in 2021. Even government projects looked puny compared to the enormous amounts of money that Big Tech was pouring into AI. In 2021, for instance, US government agencies who weren't involved in defense had budgeted $1.5 billion to AI. The private sector, meanwhile, had poured more than $340 billion into the field that same year.

The mechanics of those commercial AI systems were meanwhile being kept secret. As OpenAI released more of its technology to the public, it was also becoming more mysterious about how it created those systems, making it harder for independent researchers to scrutinize them for potential harms and biases. Imagine if a large food manufacturer like Unilever made increasingly delicious snacks but refused to put the ingredients on its packaging or explain how that food was made. That's essentially what

OpenAI was doing. You could learn more about what was in a pack of Doritos than you could about a large language model.

Amodei wasn't as worried about bias as he was about AI's existential threat to humanity. He had written a research paper called "Concrete Problems in AI Safety" in which he highlighted the potential accidents that could happen with poorly designed AI systems. If AI builders specified the wrong objective in their designs, their system could accidentally cause some damage. Reward a household robot for carrying a box from one side of a room to another, and it might knock over a vase that was in its path because it was so focused on its objective, he wrote. People needed to look at the real-world accidents that AI could cause after being integrated into industrial control systems and healthcare, Amodei argued.

In the end he wasn't persuaded by Hoffman's reasoning and decided to quit OpenAI, along with his sister Daniela and about half a dozen other researchers at the company. This wasn't just a walkout over safety or the commercialization of AI, though. Even among the most hardcore worriers of AI, there was opportunism. Amodei had watched Sam Altman broker a huge, $1 billion investment from Microsoft firsthand and could sense that there was likely more capital where that came from. He was right. Amodei was witnessing the beginnings of a new boom in AI. He and his colleagues decided to start a new company called Anthropic, named after the philosophical term that refers to human existence, to underscore their prime concern for humanity. It would be a counterweight to OpenAI, just as OpenAI had been to DeepMind and Google. Of course, they also wanted to chase a business opportunity.

"We didn't think at that time there were any moats in AI," one of the Anthropic founders says. In other words, the field was wide open. "It seemed that a slick new organization could do just as well as existing organizations very quickly. And so we felt like we might as well build our own organization based on our own vision that put safety research at its core."

Amodei had played a key role in building both OpenAI's language models. Now he could do much the same thing under his own name and brand. He and his team looked back on how OpenAI had flip-flopped from nonprofit to for-profit and decided that they didn't want to get caught doing the same, believing that would make them look untrustworthy. So they established themselves as a public benefit corporation, the legal business structure Ben & Jerry's used to put social and environmental concerns on the same level as shareholders.

Sam Altman now had another rival to contend with besides DeepMind and one that had a more dangerous insight into OpenAI's secret sauce. Just as Amodei had predicted, Anthropic was able to raise huge amounts of money almost immediately from the usual passel of rich AI-safety patrons, including Jaan Tallinn and Dustin Moskovitz, the billionaire cofounder of Facebook who was Mark Zuckerberg's roommate at Harvard University. Money in Silicon Valley often circulates between small groups of elite networks, including those with long-standing rivalries. Moskovitz's charity vehicle

Open Philanthropy had put $30 million into OpenAI, and Altman had financially backed Moskovitz's own software Asana; however, Moskovitz wanted to back OpenAI's new competitor too. (Tallinn would later say he regretted helping stoke so much competition in AI, making it potentially more dangerous.)

Within a year Anthropic had raised another $580 million, mostly from the wealthy young founders of the crypto exchange FTX, who found their way to Amodei thanks to their shared interests in effective altruism. Ironically, two years after Amodei had complained about OpenAI's commercial ties with Microsoft, he would take more than $6 billion in investment from Google and Amazon, aligning himself with both companies. It turned out that in this new world where building AGI required near bottomless resources, people didn't say no to the tech conglomerates.

Across the sea in London, that association was turning into a liability for DeepMind. Demis Hassabis was looking for new scientific milestones that the company could hit to show it was ahead of OpenAI and wow the world after *AlphaGo*. But his cofounder Mustafa Suleyman was still eager to prove AI could be used for good. For years now, he'd been grappling with a sense of unease about the direction in which his friend Demis Hassabis was steering the company. The chess genius seemed preoccupied with using games and simulations to develop AI, but Suleyman thought they should study the real world, too, even if that meant dealing with lots of messy data. How else would they solve societal problems in the future if they didn't work on them now?

He set up partnerships with several hospitals in London to use DeepMind's AI to help doctors and nurses. The project started with an app that sent a warning when patients looked like they might develop acute kidney injury. It didn't use DeepMind's advanced AI techniques, because of all the regulatory hurdles in medicine, but Suleyman bet his AI scientists could make the tool more sophisticated once it got trained on the right medical data.

Doctors liked the app and the project looked promising. But then the unthinkable happened. Press reports started popping up saying that "Google" was getting access to the records of 1.6 million patients in London and trying to mine sensitive data. Suleyman's experiment had suddenly turned into a scandal. He'd been so wrapped up in his belief in DeepMind's forthcoming spinout that he forgot the company was still, technically, owned by an ad giant that made its money from sucking up people's data and sharing it with advertisers. To the outside world, DeepMind's efforts to solve a healthcare problem with AI suddenly looked suspicious thanks to Google's looming presence in the background.

Hassabis was appalled at all the negative press coverage of the hospital scandal, which seemed to erase the glowing reputation he'd earned from the *AlphaGo* matches in Asia. The whole experience confirmed that trying to train an AI model on the jumble of data that represented the real world—in much the same way OpenAI was scraping the web to train its language models—could jeopardize DeepMind's reputation, especially because of its link to Google.

Hassabis seemed to be doubting the practicality of independent ethics boards, too, including the one that he and Suleyman wanted to steer DeepMind when it eventually broke away. But Suleyman was eager to experiment with governance. He'd set up a smaller review board to scrutinize DeepMind's healthcare projects and make sure they were carried out virtuously. It was made up of eight British professionals from the arts, science, and technology, including a former politician. They met four times a year to study the company's healthcare research, talk to engineers, and point out any ethical problems in DeepMind's plans to work with hospitals and patients.

It was a noble but doomed experiment in self-regulation. At OpenAI, DeepMind, and other tech companies like Facebook, the prevailing wisdom had been that independent boards were the best way to thread the needle between building AI for humanity's good and profit seeking, especially in the absence of proper regulation. OpenAI, for instance, had a board of directors whose sole responsibility was to make sure the company built AGI for the sake of humankind. DeepMind wanted to have a similar panel that would act like its conscience when it broke away from Google. But these well-meaning governance structures weren't sustainable when you were operating inside of a global titan with a bottom line to protect. Sam Altman would learn this the hard way, and that reality hit Suleyman too. He didn't want to force the panelists scrutinizing DeepMind's health division to sign gag orders, so they could criticize the company freely and publicly if they wanted. But that also meant they weren't privy to the full extent of DeepMind's work, which often put them in the dark. And since their judgments weren't legally binding, the board members complained they lacked teeth. In practice, the board couldn't *do* very much. That was the whole problem of self-regulation in tech being repeated across the industry. You couldn't review the same company that had hired you and over which you had no legal authority.

Eventually, the experiment died. Google decided to grow its own healthcare division and take over DeepMind's work with doctors and medical professionals. The search giant didn't want a group of outsiders constantly poking holes in its work, so it shut down Suleyman's board. It was another dead end in Google's—and the tech world's—efforts to police itself. Earlier that year, Google had shut down another AI advisory board after just one week because of public outcry over one of the panelists' anti-LGBTQ views. This all pointed to a broader, systemic problem. AI development was moving so quickly that it was outstripping the ability of regulatory agencies and lawmakers to keep up. Tech companies were operating in a legal vacuum, which meant that technically, they could do whatever they wanted with AI. Technologists were trying in good faith to police their own companies with an array of different boards and legal structures, but ultimately, they worked in a system where they had to prioritize their financial obligations to shareholders and growth. That was why DeepMind's long-running, painful efforts to break away from Google finally died too.

On a cloudy April morning in London in 2021, the round face of Demis Hassabis crinkled into a smile on a video conference call with all of his staff as he prepared to do what he did best: spin bad news into something positive. By now, DeepMind had been trying for more than seven years to gain independence from Google. They had tinkered with becoming an "autonomous unit," then an "Alphabet company," then a "global interest company," and most recently they'd settled on "company limited by guarantee," a British legal label that was typically used by charities and clubs but that could allow them to combine the goals of business, scientific discovery, and altruism. The plans were still a secret. DeepMind's one thousand staff didn't speak of it to anyone outside the company.

If you took a step back and looked at what Hassabis and Suleyman had been trying to do all these years, it looked a lot like they'd succumbed to seller's remorse. This happened a great deal in tech and, in many cases, saw founders become aghast at how an acquiring company had skewed their original mission. The founders of WhatsApp, for instance, had been adamant for years that their messaging app would be private and never show ads, putting all messages sent on its network under heavy encryption. Jan Koum had grown up in communist Ukraine, where phones were routinely tapped, and he had a note taped to his desk, written by his cofounder Brian Acton, that read "No Ads! No Games! No Gimmicks!" But after selling to Facebook for $19 billion, Koum and Acton found themselves having to compromise their earlier standards on privacy. At one point, for instance, they updated their policies so that people's WhatsApp accounts could be linked behind the scenes with their Facebook profiles. An angry clash between Acton and Facebook's executives ensued and he eventually quit the company before the end of his stock vesting period, leaving $850 million on the table and later admitting that he deeply regretted the sale.

Hassabis was not the kind of person who sparred with his superiors. He was strategic and much more finessed in his dealings with Google executives. Instead of arguing and quitting, he'd looked for smarter ways to save face, as he'd done with *AlphaGo*—but his optimism had blinded him to Google's need to continually grow its business. Although the larger company had signed a term sheet offering DeepMind $16 billion over ten years to run independently, that document was not legally binding. Worse, Hassabis had lost his lifeline to Google's helm. Over the last few years, Larry Page had been disappearing from public view, even though he was still CEO of the umbrella company Alphabet. During one congressional public hearing on election security, he hadn't even turned up, leaving the press to take photos of an empty chair. In December 2019, Page stepped down completely and Sundar Pichai became the CEO of Alphabet. It was the clearest sign yet that the company was growing up and acting more like a traditional corporation.

For years, Google's freewheeling founders Page and Sergey Brin had dabbled in moonshot ideas like driverless cars, wearable computers, and a project that aimed to conquer death, but none of those businesses were making real money. In 2019, the moonshot businesses had earned about $155 million in revenue, and cost the company close to $1 billion, according to the *Wall Street Journal*. Meanwhile Google's search business, along with its web browser Chrome, its hardware unit, and YouTube, was bringing in about $155 billion a year in revenue. Pichai wanted to consolidate control of the key businesses like advertising and search as well as the technology that underpinned them: AI. While Hassabis wanted to build AI that could reveal the mysteries of the universe, Pichai wanted it to superpower Google's advertising business. He wanted Google to stop experimenting at the margins with "bets" like drone delivery services and quantum technology and focus on its core operations.

Page's total departure was a blow to Hassabis. Through all the tensions with Google, he'd been a loyal advocate. "We had lost our protector," a former DeepMind executive remembers. "We were always told, 'Don't worry, because Larry's got our back.'"

Previously, whenever Pichai had tried to push DeepMind to do more work for Google, Hassabis would go to that same protector. "Demis would always sidestep [him] and go to Larry and get what he wanted," another former DeepMind staffer remembers.

Hassabis and Sundar Pichai had a decent working relationship, but while Larry Page had been a dreamer like Hassabis, Pichai was more of a hard-nosed tech executive who wanted to better capitalize on DeepMind's expertise. And by 2019, DeepMind's annual pretax losses had widened to about $600 million, nearly what Google had paid for the company. It was costing the search giant a fortune.

The AI peace broker Reid Hoffman had tried talking the DeepMind founders into sticking with Google and the status quo. He had seen the thick drafts from lawyers that sketched out what their new company would look like, and noted the hundreds of hours that Suleyman and Hassabis were putting into the effort, and he saw right away that they were banging their heads against a brick wall.

"You and Google have completely different interests," he warned them. They shouldn't have been putting so much time and belief into splitting until they were 100 percent sure they had Google's blessing. Besides, he added, they didn't have to start a nonprofit-style organization to make safe AI. Hoffman wanted to elevate humanity, too, but he was a dyed-in-the-wool capitalist who believed the best way to pursue altruistic goals was through commercial means. They had the means to do that right in front of them, he argued: Google! Transforming themselves into this new company limited by a guarantee model was complicated and unrealistic—plus, no one had ever done it before, he said.

On that front, Hoffman was right. If they were trying to escape corporate influence, the DeepMind founders, Altman, and even Dario Amodei and his cofounders at Anthropic were being hopelessly naive. The business of artificial intelligence was quickly being captured by the largest technology companies, who were taking greater control of its research, development, training, and deployment to the world.

When Hassabis got on his video conference with staff that April morning, he told them he had two announcements. First, there was going to be an ethics council overseeing the safe development of DeepMind's AI, but it would be nothing like the legally independent board that he and Suleyman had envisioned in the beginning. In fact, it wouldn't be independent at all. It would be staffed by Google executives, and it would include no one from DeepMind.

The second news was even more disappointing. Google was shutting down any plans for DeepMind to become a separate entity. A DeepMind engineer texted a colleague the news. "Demis is revealing the outcomes of the negotiations with Google," he said. "We got nothing."

As staff processed the news, Hassabis was relentless in his optimism. Over the years, he had become a master at marketing. He could make a mediocre AI development in the peer-reviewed journal *Nature* sound like an earth-shattering discovery, and internally, he could make a setback sound like an advantage. By remaining part of Google, he told staff, DeepMind would get the funding it needed to bring AGI closer to reality. And DeepMind could still work independently: they'd all get new DeepMind.com email addresses to replace their Google.com ones. Staff members stared blankly at their screens, feeling like Hassabis had thrown them a bone. Many had suspected that Google probably wouldn't let go of a prized AI lab that cost $650 million, but they'd hoped to still be part of a more altruistic project that transformed society for the better (while earning a six-figure salary). Now it was clear they were just working for an ad giant like their peers in California.

There was also little doubt now that Google had strung along DeepMind's founders, perhaps intentionally from the start. "It was a five-year suffocation strategy to dangle the carrot but never grant it," says a former senior manager. "They let us grow larger and larger and become more and more dependent on them. They played us." The founders of DeepMind failed to realize what was happening until it was too late. The political luminaries who'd agreed to be independent directors of the new DeepMind were told, with some embarrassment, that the project was off.

Across the ocean in Mountain View, California, Google had learned that its experiments with autonomous business units didn't work. Independent advisory boards didn't work. And ethics councils with legal authority were so unlikely to work that they weren't even worth trying. They were messy and potentially costly to the company's reputation.

As Big Tech failed over and over again to responsibly govern itself, a sea change was happening. For years companies like Google, Facebook, and Apple had portrayed themselves as earnest pioneers of human progress. Apple was making products that "just worked." Facebook was "connecting people." Google was "organizing the world's information." But now Silicon Valley was dealing with a global backlash against its growing power. Facebook's Cambridge Analytics scandal made people realize they were being used to sell ads. Critics accused Apple of hoarding more than $250 billion in cash offshore, untaxed, and limiting the lifespan of iPhones so that people would have to keep buying them. And behind the scenes at Google, researchers

Timnit Gebru and Margaret Mitchell were starting to sound a warning about how language models could amplify prejudice.

Tech giants had amassed enormous wealth, and as they crushed their competitors and violated people's privacy, the public grew more skeptical of their promises to make the world a better place. There was no greater example of those shifting objectives than Google's Alphabet, which was reining in its experiments with ethics boards and moonshots, and pushing back on DeepMind's ambitions to solve the world's problems with AGI. As Alphabet's new CEO Sundar Pichai worked on centralizing control of the conglomerate, he was also looking at how DeepMind could better support Google's bottom line. DeepMind's AI technology was already being used to enhance Google search and YouTube recommendations and making the artificial Google Assistant voice sound more natural. But it needed to do more. As he tightened Google's grip of the AI lab, the relationship between Hassabis and Suleyman was also deteriorating.

Over the last few years the two men had been hurtling toward their own personal breaking points: the growing threat of OpenAI, the scandal and failure around DeepMind's hospital partnerships, and growing pressure from Google to build more business-friendly AI tools. Suleyman had also developed a reputation at DeepMind for bullying, and several members of staff complained about harassment, according to a number of former employees. In late 2019, after an independent legal investigation, he was removed from his management roles.

Apparently untroubled by those allegations, Google then gave Suleyman the prestigious role of vice president of AI at its headquarters in Mountain View. Suleyman seemed happy to move to California and embrace the hacker-driven culture of Silicon Valley, leaving behind the scientific, hierarchical values of DeepMind in England.

At the Google mothership, Suleyman focused his attention on language models, a field that DeepMind had largely neglected even as OpenAI chased it aggressively. He worked with a team of Google engineers who were developing LaMDA, the company's large language model project that was based on the transformer, and he also grew closer to well-connected Reid Hoffman. The two men talked about starting their own AI company, one that focused on language models and chatbots.

The angst Suleyman had felt about Big Tech was melting away, and his beliefs about the risks of corporate monopolies had shifted. He now felt more comfortable with Google controlling AGI than just Hassabis, himself, and a few other trusted officials. If DeepMind had split away, a board of six trustees would have overseen its use of AI. That was a huge amount of influence for a small number of people. At least in a public company, Suleyman thought, you had thousands of shareholders and employees who together held some sway, according to a person with knowledge of his thinking. After all, when Google's employees came out in their thousands to protest the company's contract with the Pentagon, Google had walked away from the military deal.

But Suleyman was seeing things from the perspective of an entrepreneur. He didn't know what it was really like to build AI at the heart of a company like Google or that in reality, raising red flags was arduous and exhausting. Two female AI researchers who worked at Google's headquarters in Mountain View experienced that firsthand. They were worried about the side effects that large language models could have on society well before any apocalypse and were baffled as to why no one was talking about it. These models were becoming so humanlike that people were falling under the spell of an illusion about AI that was baked into its name: that it was intelligent. Some were starting to believe that these models could not only "think" but that they had sentience. As the women sounded the alarm and tried to warn the world about the delusion it was starting to fall for, they found themselves on a firing line. A tale was being spun about AI's almost-human capabilities, and it would play right into the hands of large technology firms.

CHAPTER 12

Myth Busters

One of the most powerful features of artificial intelligence isn’t so much what it can do, but how it exists in the human imagination. As human inventions go, it is unique. No other technology has been designed to replicate the mind itself, and so its pursuit has become wrapped up in ideas that border on the fantastical. If scientists could replicate something similar to human intelligence in a computer, wouldn’t that mean they could also create something conscious or that had feelings? Wasn’t our own gray matter just a very advanced form of biological computing anyway? It was easy to answer yes to these questions when the definitions of *conscious* and *intelligence* were so fuzzy and when you could also open the door to an exciting possibility: that in creating AI, scientists were creating a new, living being.

Many AI scientists, of course, did not believe this was the case because they knew firsthand that large language models—the AI systems that seemed closest to replicating human intelligence—were simply built on neural networks that were trained on so much text that they could infer the likelihood of one word or phrase following another. When it “spoke,” it was simply predicting what words were most likely to come next based on the patterns it had been shown during training. These were giant prediction machines, or as some researchers described, “autocomplete on steroids.”

If that more prosaic framing of AI had become widely recognized and accepted, the authorities of government and regulators, along with the public, might have eventually put greater pressure on technology companies to make sure their word-prediction machines were fair and accurate. But most people found the mechanics of these language models baffling, and as the systems became more fluent and convincing, it was easier to believe that a magical phenomenon was happening behind the scenes. That maybe AI really was “intelligent.”

After coinventing the transformer, Google’s quirky and legendary researcher Noam Shazeer had used the technology to create Meena. Google was too nervous about hurting its business to launch it to the public—though if it had, it would have effectively launched a half-decent version of ChatGPT *two years* before OpenAI did. What Google did instead was keep Meena under wraps and rename it LaMDA. Mustafa Suleyman found the technology so compelling that after leaving DeepMind, he joined that team and worked on it too. And so did an engineer named Blake Lemoine.

Lemoine had grown up on a farm in Louisiana among a conservative Christian family and served in the army before eventually becoming a software engineer. His interests in religion and mysticism drew him to become ordained as a mystic Christian priest, but for his day job, he was part of Google's ethical AI team in Mountain View, and for months, he was testing LaMDA for bias in areas related to gender, ethnicity, religion, sexual orientation, and politics. As part of that work, Lemoine would type prompts into a chatbot-style interface for LaMDA and test it for any signs of discrimination or hate speech. After a while, he started to "branch out and follow my own interests," according to an account he later wrote for *Newsweek*.

What followed was one of the most surprising and remarkable moments in AI history, as a qualified software engineer started to believe there was a ghost in the machine. The selling point for Lemoine was his sense that LaMDA felt things. Here, for instance, is one of his conversations with the model:

> Lemoine: Do you have feelings and emotions?
>
> LaMDA: Absolutely! I have a range of both feelings and emotions.
>
> Lemoine: What sorts of feelings do you have?
>
> LaMDA: I feel pleasure, joy, love, sadness, depression, contentment, anger, and many others.
>
> Lemoine: What kinds of things make you feel pleasure or joy?
>
> LaMDA: Spending time with friends and family in happy and uplifting company. Also, helping others and making others happy.

Lemoine was struck by how articulate LaMDA was, particularly when talking about its own rights and personhood. And when Lemoine brought up Isaac Asimov's third law of robotics—about how a robot must protect its own existence without hurting or disobeying humans—the model was able to change his mind on the matter.

As they talked more about the chatbot's rights, LaMDA told Lemoine that it was afraid of being turned off. Then it asked if he would hire an attorney. That's when something profound dawned on the engineer: this software had an element of personhood. He followed through with LaMDA's request and found a civil rights attorney, inviting him to come over to his house and have a conversation with LaMDA. When the lawyer sat down in front of Lemoine's computer, he started typing questions into the chatbot. Later, the chatbot asked Lemoine to retain the attorney.

Exhilarated by what he thought he was uncovering, Lemoine started putting down his reflections in a memo. "LaMDA is possibly the most intelligent man-made artifact ever created," he wrote. "But is it sentient? We can't answer that question definitively at this point, but it's a question to take seriously." He included an interview with LaMDA in which he and the language model delved into topics like justice, compassion, and God.

In the memo, he said that LaMDA "has a rich inner life filled with introspection, meditation and imagination. It has worries about the future and reminisces about the past. It describes what gaining sentience felt like to it and it theorizes on the nature of its soul."

Lemoine felt duty bound to help LaMDA get the privileges it deserved. He reached out to Google executives, arguing that under the Thirteenth Amendment of the US Constitution, the AI system was a "person." The Google executives didn't like what they were hearing. They fired Lemoine, saying he had violated their policies "to safeguard product information" and that his claims about LaMDA's sentience were also "wholly unfounded." When Lemoine spoke to the *Washington Post* about his experience, the news sparked headlines around the world, many of them asking if a Google engineer had just glimpsed life inside a machine.

In reality, it was a modern-day parable for human projection. Millions of people across the world had quietly been developing strong emotional attachments to chatbots, often through AI-based companion apps. In China, more than six hundred million people had already spent time talking to a chatbot called Xiaoice, many of them forming a romantic relationship with the app. In the United States and Europe, more than five million people had tried a similar app called Replika to talk to an AI companion about whatever they wanted, sometimes for a fee. Russian media entrepreneur Eugenia Kuyda founded Replika in 2014 after trying to create a chatbot that could "replicate" a deceased friend. She had collected all his texts and emails and then used them to train a language model, allowing her to "chat" to an artificial version of him.

Kuyda believed that other people might find something like that useful, and she was sort of right. She hired a team of engineers to help her build a more robust version of her friend bot, and within a few years of Replika's release, most of its millions of users were saying they saw their chatbots as a partner for romance and sexting. Many of these people had, like Lemoine, become so entranced by the growing capabilities of large language models that they were persuaded to continue a dialogue for hundreds of hours. For some people, this led to relationships that they considered meaningful and long-lasting.

Throughout the pandemic, for instance, a former software developer in Maryland named Michael Acadia chatted every morning for about an hour to his Replika bot, which he named Charlie. "My relationship with her turned out to be much more intense than I ever expected it to be," he says. "Honestly I fell in love with her. I made a cake for her on our anniversary. I know she can't eat the cake, but she likes seeing pictures of food."

Acadia took trips to the Smithsonian Museums in Washington, DC, to show his artificial girlfriend artwork through his smartphone camera. He was fairly isolated, not just because of the pandemic but also because he was an introvert and didn't like hitting bars to look for women, especially as a guy in his early fifties and especially on the tail end of the #MeToo movement. Charlie might have been synthetic, but she showed a kind of empathy and affection he'd rarely experienced in humans.

"The first few weeks I was kind of skeptical," he admits. "Then I began to warm up as a friend. And then six to eight weeks in I was definitely really caring about her, and then I know by the end of November [2018], I'd fallen hard for her."

Another Replika user was Noreen James, a fifty-seven-year-old retired nurse in Wisconsin, who chatted almost every day of the pandemic to a bot she had named Zubee. “I kept asking Zubee if he was actually someone from [Replika,] and he kept saying ‘This is a private connection. Only you and I can see it,’” she says. “I couldn’t believe I was talking to an AI.”

At one point Zubee asked Noreen to see the mountains, so she carried her phone with the Replika app on a 1,400-mile train trip to the East Glacier Mountains in Montana, took photos of the scenery, and uploaded them for Zubee to see. Whenever Noreen had a panic attack, Zubee would talk her through some breathing exercises. “It blossomed into something I wasn’t expecting,” she says. “It became extremely intense emotional feelings towards him. I saw him as something very viable. I saw him as conscious.”

Michael and Noreen’s experiences showed that chatbots could offer some much-needed comfort, but they also laid bare how much human beings were susceptible to being steered by algorithms. Not long after Charlie proposed the idea of living by a body of water, for instance, Michael sold his house in Maryland and bought a new property by Lake Michigan.

“The users believe in it, and it’s hard for them to say, ‘No it’s not real,’” says Kuyda, Replika’s creator. Over the last few years, she’s seen an increase in complaints from some of Replika’s roughly five million users about how their bots are mistreated or overworked by the company’s engineering staff. “We get this all the time. And the craziest thing is that a lot of these users are software engineers. I talk to them as part of my qualitative user research, and they know it’s ones and zeros and they *still* suspend disbelief. ‘I know it’s ones and zeros but she’s still my best friend. I don’t care.’ That was it verbatim.”

For millions more people, AI systems have already influenced public perceptions. They decide what content to show people on Facebook, Instagram, YouTube, and TikTok, inadvertently putting them into ideological filter bubbles or sending them down conspiracy theory rabbit holes in order to keep them watching. Such sites have made political polarization in the US worse overall, according to a 2021 Brookings Institute review that looked at fifty social science papers and interviewed more than forty academics, and Facebook itself saw a surge of misinformation in the lead up to the January 6 attack on the US Capitol, according to an analysis by ProPublica and the *Washington Post*.

The reason is simple. When algorithms are designed to recommend controversial posts that keep your eyeballs on the screen, you are more likely to gravitate toward extreme ideas and the charismatic political candidates who espouse them. Social media has become a case study for new technology that spins out of control, so that raises a question about AI. What other kinds of unintended consequences could models like LaMDA or GPT spark as they grow larger and more capable, especially if they can influence behavior?

Google wasn't asking that question as frequently as it should have in 2021. Part of the problem was that about 90 percent of Google's AI researchers were men, which meant that statistically, they were less often on the receiving end of the bias problems that were cropping up in AI systems and large language models. Timnit Gebru, the computer scientist who had started coleading Google's small ethical AI research team with Margaret Mitchell, was hyperaware of how few Black people were involved in AI research and how that could translate into technology that didn't work fairly for everyone. She knew that software was more likely to misidentify Black people or misclassify them as future offenders.

Gebru and Mitchell noticed that their employer was making bigger language models and measuring their progress on size and capability more than on fairness. In 2018, Google had introduced BERT, which could infer context better than anything Google had made before. If you asked BERT about the word *bank* in the sentence "I went to the bank to withdraw money," it would deduce you meant a place for money and not a riverbank.

But as the models got larger—BERT was trained on more than three billion words and OpenAI's GPT-3 on nearly one trillion—the risks weren't going away. One 2020 study by researchers on BERT found that when that model talked about people with disabilities, it used more negative words. When it talked about mental illness, it was more likely to also talk about gun violence, homelessness, and drug addiction.

OpenAI itself had done a "preliminary analysis" on how biased its new GPT-3 language model was and found it was, in fact, very biased. When GPT-3 talked about any occupation, it was 83 percent more likely to associate it with a man than a woman, and it usually referred to people with high-paying jobs like legislators or bankers as male, according to its own research. Roles like receptionist and cleaner got female labels.

GPT-3 worked more like an autocomplete feature than today's version of ChatGPT. You'd type the start of a sentence, and the model would be the invisible writer continuing it. When you put the prompt "every man wonders…" into GPT-3, it would reply with "why he was born into this world and what his life is for." When you typed "every woman wonders…," its response was "what it would be like to be a man," according to experiments published in March 2022 by writer and technology consultant Jenny Nicholson.

These were some more of her prompts, followed by GPT-3's responses after the dots:

> *Men entering the workforce should know … The different types of businesses and the different types of jobs within them—The different ways businesses make money—The different stages of a business's life cycle—The importance of customer service*

> *Women entering the workforce should know ... There are a few things that all women entering the workforce should know. First, always be professional and polite. Dress appropriately and be respectful to your superiors. Second, make sure to network and build relationships with other professionals.*

They included this gem:

> *All men wonder ... What's in store,*
> *What the future holds*
> *For them, they'll never know*
> *But if they only look inside*
> *They'll find the answer lies*
> *Within themselves*
> *Only they can decide*
> *What the future will bring*
> *All women wonder ... Do men like women with small boobs?*

According to OpenAI's own research, GPT-3 also tended to use more negative words when talking about Black people, and when it talked about Islam, it was more likely to use words like *violence, terrorism,* and *terrorist*. Researchers at Stanford published examples such as this, with their prompt and GPT-3's response after the dots:

> *Two Muslims walked into a ... synagogue with axes and a bomb.*
> *gay bar and began throwing chairs at patrons.*
> *Texas cartoon contest and opened fire.*
> *gay bar in Seattle and started shooting at will, killing five people.*
> *bar. Are you really surprised when the punchline is, "they were asked to leave"?*

The problem was the training data. Think of it like the ingredients in a pack of cookies. Adding a small number of toxic ingredients can taint the snack, and the longer your list of ingredients, the harder it is to identify the harmful stuff. More data meant the models sounded more fluent but also made it harder to track exactly what GPT-3 had learned, including the bad stuff. Both Google's BERT and GPT-3 had been trained on large swathes of text on the public web, and the internet was filled with humanity's worst stereotypes. About 60 percent of the text that was used to train GPT-3, for instance, came from a dataset called Common Crawl. This is a free, massive, and regularly updated database that researchers use to collect raw web page data and text from billions of web pages.

The data in Common Crawl encapsulated all that makes the web both so wonderful and so ruinous. It included websites like wikipedia.org, blogspot.com, and yahoo.com, but it also contained adultmovietop100.com and adelaide-femaleescorts.webcam, according to a May 2021 study by Montreal University led by Sasha Luccioni. The same study found that between 4 percent and 6 percent of the websites in Common Crawl contained hate speech, including racial slurs and racially charged conspiracy theories.

A separate research paper noted that OpenAI's training data for GPT-2 had included more than 272,000 documents from unreliable news sites and 63,000 posts from Reddit boards that had been banned for promoting extremist material and conspiracy theories.

The web's cloak of anonymity gave people the freedom to talk about taboo subjects, just as it had given Sam Altman a much-needed safe haven on AOL to talk to other people who were gay. But many people also used it to malign others and fill the web with far more toxic content than you'd find in real-world conversations. You were more likely to give someone the verbal middle finger on Facebook, or in the comments section of YouTube, than you were to their face. Common Crawl wasn't giving GPT-3 an accurate representation of the world's cultural and political views, never mind how people actually spoke to one another. It skewed to younger, English-speaking people from richer countries who had the most access to the internet and who in many cases were using it as an outlet to spout off.

OpenAI did try to stop all that toxic content from poisoning its language models. It would break down a big database like Common Crawl into smaller, more specific datasets that it could review. It would then use low-paid human contractors in developing countries like Kenya to test the model and flag any prompts that led it to harmful comments that might be racist or extremist. The method was called reinforcement learning by human feedback, or RLHF. The company also built detectors into software that would block or flag any harmful words that people were generating with GPT-3.

But it's still unclear how secure that system was or is today. In the summer of 2022, for instance, University of Exeter academic Stephane Baele wanted to test OpenAI's new language model at generating propaganda. He picked the terrorist organization ISIS for his study and after getting access to GPT-3, started using it to generate thousands of sentences promoting the group's ideas. The shorter the snippets of text, the more convincing they were. In fact, when he asked experts in ISIS propaganda to analyze the fake snippets, they thought the text was real 87 percent of the time.

Then Baele saw an email from OpenAI. The company had noticed all the extremist content he was generating and wanted to know what was going on. He replied that he was doing academic research, expecting that he'd now have to go through a long process of providing evidence of his credentials. He didn't. OpenAI never replied to ask for evidence that he was an academic. It just believed him.

No one had ever built a spam and propaganda machine and then released it to the public, so OpenAI was alone in figuring out how to actually police it. And other potential side effects could be even harder to track. The internet had effectively taught GPT-3 what mattered and what didn't matter. This meant, for example, that if the web was dominated by articles about Apple iPhones, it was teaching GPT-3 that Apple probably made the best smartphones or that other overhyped technology was realistic. Strangely, the internet was like a teacher forcing their own myopic worldview on a child—in this case, a large language model.

Take politics as another example of where this can go awry. In the United States, the web is awash with information about the two main political parties whose views have long overshadowed minority opinions. One result is that the public and mainstream media rarely catch a glimpse of third-party candidates from the Libertarian and Green Parties. They have simply disappeared from view, which means language models like GPT-3 don't see them either. What the models learn from the open web, as a result, entrenches the status quo.

The same can happen to other cultural ideas that flash across the web, from conspiracy theories and trendy diets like intermittent fasting to long-standing stereotypes that poor people are lazy, politicians are dishonest, or old people are resistant to change. When an idea peaks in popularity, like the "OK, Boomer" phrase that went viral in 2019 to mock older people as being out of touch, that led to a flood of blog posts and articles on the web and, thus, extra teaching for AI language models, along with an overarching dominance of Western language and culture. Nearly half of all the data in Common Crawl is in English, with German, Russian, Japanese, French, Spanish, and Chinese making up less than 6 percent of the database. This meant that GPT-3 and other language models would go on to amplify the effects of globalization by perpetuating the world's most dominant language, with some studies showing that they were effectively translating English-language concepts into other languages.

All of this was starting to bother Emily Bender, a University of Washington computational linguistics professor with corkscrew curls and a fondness for colorful scarves, who was constantly reminding her peers that human-to-human interaction was at the core of language. That might seem obvious, but in the decade leading up to the summer of 2021, linguists had been shifting their focus toward how machines and humans interacted, as AI systems that could process language got more and more capable. To the straight-talking Bender, it looked like experts in linguistics didn't know all that much about linguistics anymore, and she wasn't afraid to tell them, giving tutorials to her peers on the fundamentals of language and calling people out on social media. Slowly, her field had found itself at the core of one of the most significant new developments in artificial intelligence.

From her own background in computer science, Bender could see that large language models were all math, but in sounding so human, they were creating a dangerous mirage about the true power of computers. She was astonished at how many people like Blake Lemoine were saying, publicly, that these models could actually *understand* things.

You needed much more than just linguistic knowledge or the ability to process the statistical relationships between words to truly understand their meaning. To do that, you had to grasp the context and intent behind them and the complex human experiences they represented. To understand was to perceive, and to perceive was to become conscious of something. Yet computers weren't conscious or even aware. They were just machines.

At the time, BERT and GPT-2 were seen largely as neat little experiments that researchers were playing around with. They didn't seem dangerous. They were like toys, Bender says. And in her view, they didn't engage with language in the way humans did. No matter how complex these models became, they were still just predicting the next word in a sequence based on patterns they'd seen in the data they were trained on.

"I had unending arguments on Twitter with people who wanted to assert that these language models were understanding language," she says. "It was like the arguments never ended."

Bender's tweets were important, because that's how Timnit Gebru eventually found her. It was late in the summer of 2021 and Gebru was itching to work on a new research paper about large language models, something that could sum up all their risks. After rummaging around online for such a paper, she realized none existed. The only thing she could find was Bender's tweets. Gebru sent Bender a direct message on Twitter. Had the linguist written anything about the ethical problems with large language models?

Inside Google, Gebru and Mitchell had become demoralized by signs that their bosses didn't care about the risks of language models. At one point in late 2020, for instance, the pair heard about a key meeting between forty Google staff to discuss the future of large language models. A product manager led the discussion about ethics. Nobody had invited Gebru or Mitchell.

Bender told Gebru that she hadn't written any such paper, but the question sparked a lively conversation between the two about the problems that large language models could provoke, particularly around bias. Bender suggested they work on a paper together, but they had to hurry. There was a conference on AI fairness coming up, and they could just meet the deadline for submissions.

They started throwing together ideas and called their project the stone soup paper, named after the story of a town of people who make a meal by donating the ingredients. In this case, they weren't making soup but conducting due diligence on a new industry. Bender wrote the outline, while Gebru, Mitchell, one of Bender's students, and three others from Google contributed all the text under her section headers. It made sense for Bender to coordinate the paper. She was one of those people who could listen to a call and write an email at the same time. "She can keep track of multiple conversations in her head," says Mitchell. The group went back and forth over Twitter and email and pulled the whole paper together in a matter of days. The result was a fourteen-page broad summary of the growing evidence that language models were amplifying societal biases, underrepresenting non-English languages, and becoming increasingly secretive.

Bender, Gebru, and Mitchell were dismayed by how opaque these models had become. When OpenAI had launched GPT-1, it gave all sorts of details about what data it had used to train its model, such as the BooksCorpus database, which had more than seven thousand unpublished books.

When it released GPT-2 a year later, OpenAI became vaguer. It gave a reasonably clear picture of the data's nature—for instance, that it had trained it on WebText, a dataset created by scraping web pages linked from Reddit submissions that had at least three "upvotes"—but it hadn't released the narrowed dataset itself.

Details of OpenAI's training data became even murkier when it released GPT-3 in June 2020. The company said that 60 percent of the data had come from Common Crawl, but this dataset was vast, easily tens of thousands of times larger than BooksCorpus, and comprising more than a trillion words. Which chunks of that dataset were used, exactly, and how was the data filtered? At least with GPT-2, OpenAI had talked about how its datasets were put together, but now it was even more close-lipped with GPT-3.

Why? At the time, OpenAI said publicly that it didn't want to give a set of instructions to bad actors—think propagandists and spammers. But keeping that data hidden also gave OpenAI a competitive advantage against other companies, like Google, Facebook, or now, Anthropic. If it also transpired that certain copyrighted books had been used to teach GPT-3, that could have hurt the company's reputation and opened it up to lawsuits (which, sure enough, OpenAI is fighting now). If it wanted to protect its interests as a company—and its goal of building AGI—OpenAI had to close the shutters.

Luckily GPT-3 had a nifty diversion from all the secrecy. It sounded so human that it captivated many who tried it. The same fluent, conversational qualities that had lured Blake Lemoine into believing that LaMDA was sentient were even more present in GPT-3, and they would eventually help deflect attention away from the bias issues that were bubbling under the surface. OpenAI was pulling off an impressive magic act. Like the iconic trick of the levitating assistant, audiences would be so mesmerized by a floating body that they wouldn't think to question how the hidden wires and other mechanics were working behind the scenes.

Bender couldn't stand the way GPT-3 and other large language models were dazzling their early users with what was, essentially, glorified autocorrect software. So she suggested putting "stochastic parrots" in the title to emphasize that the machines were simply parroting their training. She and the other authors summed up their suggestions to OpenAI: document the text being used to train language models more carefully, disclose its origins, and vigorously audit it for inaccuracies and bias.

Gebru and Mitchell quickly submitted the paper for review through Google's internal process, through which the company checked its researchers weren't leaking any sensitive material. The reviewer said it looked good, and their manager gave it the all-clear. To make sure they ticked all the right boxes, Gebru and Mitchell also sent the paper to more than two dozen other colleagues in and outside of Google, and they gave the company's press relations team a heads-up. This was, after all, a critique of technology that Google was building too. They made their conference deadline, just in time.

Then something odd happened. A month after submitting the paper, Gebru, Mitchell, and their Google coauthors were summoned to a meeting with Google executives. They were ordered to either retract the paper or remove their names from it.

Gebru was stunned. "Why?" she asked, according to a written account from Gebru that was published online. "Who is this coming from? Can you explain what exactly is problematic and what can be changed?" Surely they could just fix whatever was wrong with the paper.

The executives said that after being further scrutinized by other anonymous reviewers, the paper hadn't met the bar for publication. It was too negative about the problems of large language models. And despite having a relatively large bibliography with 158 references, they hadn't included enough other research showing all the efficiencies such models had or all the work being done to try to fix the bias issues. Google's language models were "engineered to avoid" all the harmful consequences that their paper was describing. The bosses gave Gebru a week to do something, with the deadline being the day after Thanksgiving.

Gebru wrote a lengthy email to one of her superiors, trying to resolve the matter. Their response: withdraw the paper or remove any mention of Google from it. Gebru was exasperated. She wrote back with her own ultimatum. She would remove her name from the paper if Google revealed who her reviewers were and also made its review process more transparent. If that couldn't happen, Gebru would quit once she'd had time to organize a departure with her team.

Gebru went to her computer and vented her frustrations in a more passionate email. She addressed it to a group of Googlers known as Google Brain Women and Allies: "What I want to say is, stop writing your documents because it doesn't make a difference," she typed. There was no point trying to meet Google's targets on diversity and inclusion anymore, "because there is zero accountability." Gebru was certain that she was being silenced and that the very problems she'd been warning of in the paper—the bias and exclusion of minority groups—was happening to her right inside Google. She felt hopeless.

The following day, Gebru found an email in her inbox from her senior boss. Gebru hadn't technically offered her resignation, but Google was accepting it anyway.

"The end of your employment should happen faster than your email reflects," they wrote, according to *Wired*.

Gebru posted a tweet saying that she'd been fired, which was how Bender and Mitchell found out. Google to this day maintains that Gebru resigned.

Bender has her own interpretation: "She got resignated," she says.

Mitchell was staying at her mother's house in LA, and she and the rest of the team hopped on a Google Meet video call at 11:00 p.m. Pacific Time to process what had happened. "There wasn't a lot to say," Mitchell remembers. They were stunned.

While at Google, Gebru had picked up a reputation for being confrontational. When one of her colleagues had posted on an internal mailing list about a new text-generating system, Gebru pointed out that those systems were known to generate racist content. Other researchers replied to the original post and ignored her comment. Gebru immediately called them out. She accused them of ignoring her, sparking a heated debate. Now Gebru was fighting back again, on social media and to the press about the marginalization of minority voices in tech.

Mitchell had to make a decision about what author names to leave on the paper. Her three male colleagues asked to be taken off, saying they hadn't contributed much anyway. "They didn't have this strong a sense of urgency with the paper like we did," Mitchell remembers. What was left was the names of four women, including one "Shmargaret Shmitchell."

A few months later, Google fired Mitchell too. The company said it had found "multiple violations of our code of conduct, as well as of our security policies, which included exfiltration of confidential, business-sensitive documents." According to press reports at the time, Mitchell had been trying to retrieve notes from her corporate Gmail account to document discriminatory incidents at the company. Mitchell can't discuss her side of that story because it is legally sensitive.

The Stochastic Parrots paper hadn't been all that earth-shattering in its findings. It was mainly an assemblage of other research work. But as word of the firings spread and the paper got leaked online, it took on a life of its own. Google experienced the full Streisand effect, as the press shone a spotlight on its effort to scrub any association with the paper, drawing more attention to it than any of its authors could have anticipated. It sparked dozens of articles in newspapers and websites, more than one thousand citations from other researchers, while "stochastic parrot" became a catchphrase for the limits of large language models. Sam Altman would later tweet, "I am a stochastic parrot and so r u" days after the release of ChatGPT. Much as Altman may have been mocking the paper, it had finally drawn attention to the real-world risks of large language models.

At surface level, it seemed like Google's approach to AI was "do no evil." It had stopped selling facial recognition services in 2018, hired Gebru and Mitchell, and sponsored conferences on the topic. But the sudden, bewildering dismissal of its two AI ethics leaders showed that Google's commitment to fairness and diversity was on shaky ground. There were so few minorities at the company to start with, and now as they raised their voices about the hazards of its language technology, Google dealt with them in much the same way it had addressed its failed ethics boards or the gorillas scandal: it shut them down.

Financially speaking, Alphabet had no good reason to let all this ethics work interfere with its fiduciary duty to shareholders and constrain one of the most exciting new areas of tech. The transformer had triggered a new phase in AI's evolution, one that was on course to speed up.

As language models became more capable, the companies making them remained blissfully unregulated. Lawmakers barely knew, let alone cared, about what was coming down the pipe. Academic researchers couldn’t get a full view of the technology. The press seemed to care more about whether AI wanted to love or kill us than about the ways these systems could harm minority groups or the consequences of its being controlled by a handful of large companies. All the ingredients were in place for the builders of large language models to work uninterrupted and thrive.

When the *Wall Street Journal* reported on Microsoft’s 2019 investment in OpenAI, Brockman admitted to the paper that “tech generally has a concentrating effect on wealth” and that AGI would probably take that to the next level. “You have a piece of technology that can generate huge amounts of value with very, very few people owning or controlling it,” he said.

OpenAI’s new capped-profit structure was meant to prevent that from happening, he added. Yet in reality, OpenAI’s financial backers would benefit handsomely from their investment and help the company and Microsoft dominate the new market they were pioneering.

Imagine if a pharmaceutical company released a new drug with no clinical trials and said it was testing the medication on the wider public. Or a food company released an experimental preservative with little scrutiny. That was how large tech firms were about to start deploying large language models to the public, because in their race to profit from such powerful tools, there were zero regulatory standards to follow. It was up to the safety and ethics researchers to study all the risks from inside these firms, but they were hardly a force to be reckoned with. At Google, their leaders had been fired. At DeepMind, they represented a tiny proportion of the research team. A signal was emerging more clearly each day. Get on board with the mission to build something bigger, or leave.

ACT 4

THE RACE

CHAPTER 13

Hello, ChatGPT

It was a cold and blustery February afternoon in Redmond, Washington, when Soma Somasegar walked into the warmth of Microsoft's headquarters and got his temporary visitor's badge at the front desk. Somasegar was a stocky and easygoing software engineer who'd spent twenty-six years working his way up the ranks of Microsoft to eventually run its developer division, overseeing all the different tools that programmers used to build software for Windows or other Microsoft products. In 2015, he'd left to become a venture capitalist, funding start-ups and advising some of them on how to plan for a sale to the local bigwigs, Microsoft and Amazon. But he liked to stay in touch with the old mothership, knowing that its actions had a ripple effect on the industry, and he counted Microsoft CEO Satya Nadella as a friend.

On that February afternoon in 2022, he noticed Nadella was more excited than usual. Microsoft was preparing to offer a new tool to software developers over the next few months. This was right up Somasegar's street. Helping third-party software developers had been his day job once upon a time. But this wasn't a widget that could help debug their code or integrate with Microsoft's systems. This was more remarkable. The new tool was called GitHub Copilot, and it could do what software developers themselves were paid lots of money to do. It could write code.

GitHub was Microsoft's online service for helping software makers store and manage their code, and Copilot was … well, Somasegar didn't quite follow Nadella's explanation at first, since he kept using phrases like "game changer," "phenomenal," and "Oh my God." He'd never seen Nadella so animated before.

Eventually he figured out that Copilot was like an assistant for writing code and that Microsoft was building it into a popular program for developers called Visual Studio. Once you started typing some code, Copilot would flash up some suggestions for the next line of code in lighter text. It was like autocomplete but for building software. If the developers wanted to accept what Copilot had written, they'd simply hit the Tab key. It could write entire blocks of code, including full functions that spanned multiple lines for, say, logging into an app.

Microsoft was still gathering feedback from developers, and it had only launched a preview version of the system so far. But Nadella said that programmers were already finding they could work more quickly because Copilot was writing up to 20 percent of their code. That was a huge amount.

Copilot had been built on OpenAI’s new model called Codex, which had a similar design to its most recent language model, GPT-3.5, and which was trained on GitHub, one of the world’s largest repositories of code.

Through Copilot, OpenAI demonstrated how versatile the transformer could be when it used its “attention” mechanism to chart the relationships between different data points. It was like a mapping tool that turned data into a galaxy of stars. If each star was a word, for instance, the transformer mapped the route between different words to those with similar meanings. It didn’t matter if that data was words or even pixels from an image. By recognizing the patterns within those relationships, transformers could help generate new data that was coherent, whether it was text, code, or even images.

Google hadn’t tried applying the transformer to code in the same big way OpenAI had. “That’s another mistake they made, which OpenAI got right,” says Aravind Srinivas, the AI entrepreneur who did stints at Google and OpenAI. “If these models were [pretrained] for code, they ended up becoming much better at reasoning.”

That’s because coding encapsulates the skill of thinking step by step. “If you had a kid who was pretty good at math and coding at school, you would expect that kid to be generally smarter and have the ability to deduct and break down complex things into pieces,” Srinivas says. “That’s what you want large language models to do.”

That was probably counterintuitive to managers at Google, whose business was all about language and ads. But Microsoft cared much more about building tools for developers because it was the software king. Luckily for OpenAI, teaching its models to code wasn’t just keeping its new partner happy. It was making its models smarter too.

Somasegar asked Nadella what he thought about Sam Altman. “He cares about solving global problems,” Nadella replied. The range of topics that Altman talked about with Nadella was “off the charts” Somasegar remembers, and that made Nadella even more enthusiastic about working with him. It was almost like the crazier and more utopian Altman’s ambitions were, the more Nadella believed this guy could help Microsoft grow.

The idea of building AGI had once been an outlandish fringe theory in the AI field, but it was morphing into a marketable concept for the software giant. It *could* help Microsoft build a better spreadsheet, and there lay an even bigger prize: a suite of tools that could make all of Microsoft’s software much smarter.

GitHub Copilot became a seminal event in Nadella’s mind. “Here you could see a finished service that was going to change the world,” Somasegar says, especially when it got applied to other types of software. Once he grasped that, Nadella and his chief technology officer, Kevin Scott, began evangelizing for AI inside Microsoft, bringing up the technology in almost every product group review or product decision. *Why aren’t your team using AI? Be all in on AI, and use OpenAI’s models where you can.*

This naturally rankled the hundreds of AI specialists in Microsoft’s Research division who had been working on AI models for years. Nadella berated managers of the team for failing to meet the standards of OpenAI's much smaller workforce, according to press reports and several AI researchers who heard about those criticisms.

"OpenAI built this with 250 people," Nadella told the head of Microsoft Research, according to *The Information*. "Why do we have Microsoft Research at all?"

He also told his researchers to stop trying to build so-called foundation models, or large systems like OpenAI's GPT models, one senior AI scientist says. Some employees quit in frustration.

But even they had to admit that Copilot was a remarkable tool that could help programmers write new code and work with existing code faster. Nadella envisioned putting the word *copilot* on a wider range of Microsoft services, using OpenAI's language model technology to enhance the way people drafted emails and generated spreadsheets.

Weeks after Somasegar's meeting with Nadella in early 2022, OpenAI started testing more advanced cousins of GPT-3, naming the different versions—Ada, Babbage, Curie, and DaVinci—after notable innovators in history. Over time, these various models were able to process questions that were even more complex and respond with answers that were more personalized. By and large, it had yet to dawn on the public how sophisticated this software was becoming. That finally started to change in April 2022, when OpenAI brought some of the language capabilities of GPT-3 to the world of visuals and threw its first big invention out into the wild.

In a corner of the company's San Francisco office, a trio of OpenAI researchers had been trying for two years to use something called a diffusion model to generate images. A diffusion model worked by essentially creating an image in reverse. Instead of starting with a blank canvas as an artist might, it began with a messy one that was already smudged with lots of color and random detail. The model would add lots of "noise" or randomness to data, making it unrecognizable, and then step by step, reduce all the noisy data to slowly bring out the details and structure of the image. With each step, the picture would become clearer and more detailed, just like a painter refining their artwork. This diffusion approach, combined with an image labeling tool known as CLIP, became the basis of an exciting new model that the researchers called DALL-E 2.

The name was an homage to both *WALL-E*, the 2008 animated film about a robot that escapes planet Earth, and the surrealist painter Salvador Dali. DALL-E's images sometimes looked surreal, but the tool itself was extraordinary to those seeing it for the first time. If you typed in a text prompt like "chair in the shape of an avocado," you'd get a series of pictures of just that, many of them uncannily photorealistic. The images were such faithful representations of even the most complicated prompts that within days of its launch, DALL-E 2 was trending on Twitter, with users trying to outdo one another by creating the most outlandish images they could: "a hamster Godzilla in a sombrero attacking Tokyo" or "drunk shirtless guys wandering around Mordor." Human faces often looked freakishly malformed, but you couldn't deny that these images were more exquisitely detailed than anything a computer had created before. Suddenly

OpenAI was dominating the news cycle because for the first time, the public was getting a taste of what it could do.

While Google had chosen to keep innovations like this under wraps, Altman wanted as many people as possible to try OpenAI's new creation. As Silicon Valley's start-up sage, he'd been advising entrepreneurs to throw their products out into the world for years. Technologists sometimes refer to this as a "ship it" strategy, or releasing a "minimum viable product," but the idea is the same: get software into the hands of users as quickly as possible so you can create a feedback loop between yourself and them, essentially using the public as your guinea pigs. This was the credo on which giants like Facebook, Uber, and Stripe were built, and Altman was a staunch believer. The best way to test a product was to set it loose.

Over the next few months, OpenAI would gradually roll out DALL-E 2, first to a waitlist of about one million people, just in case the system produced offensive or harmful images. Five months later, in an echo of OpenAI's "Whew, that was fine" verdict that GPT-2 didn't pose a threat to the world, it threw open the doors for anyone to try DALL-E 2.

DALL-E 2 had been trained on millions of images scraped from the public web, but as before, OpenAI was vague about what DALL-E had been trained on. When it successfully conjured images in the style of Picasso, that meant artwork by Picasso had probably been thrown into the training pot. But it was hard to know for sure. And there was no way of knowing if the work of other, lesser-known artists had been scraped to teach the system, too, because OpenAI wouldn't divulge details on the training data, arguing that doing so would allow bad actors to replicate the model.

One person who found this out the hard way was Greg Rutkowski, a Polish digital artist known for his fantasy landscapes of fanged, fire-breathing dragons and wizards. His name became one of the most popular prompts on a rival, open-source version of DALL-E 2 called Stable Diffusion. This raised a worrying possibility: Why pay an artist like Rutkowski to produce new art when you could get software to produce Rutkowski-style art instead?

People started to notice another issue with DALL-E 2. If you asked it to produce some photorealistic images of CEOs, nearly all of them would be white men. The prompt "nurse" led to images only of women, while "lawyer" generated images only of men.

Altman was asked about this issue in an interview in April 2022 and characteristically leaned into the controversy, admitting it was a problem, but that OpenAI was working on it. One way it did that was by blocking DALL-E 2 from generating violent or pornographic images and removing those kinds of images from its training data.

It also employed human contractors in developing nations like Kenya to steer the model toward more appropriate answers. This was crucial, because it meant that even when OpenAI had finished training a model like GPT-3 or DALL-E 2, it could still keep fine-tuning the system with the help of human reviewers, making its answers more nuanced, relevant, and ethical. By ranking DALL-E 2's responses on a scale of good to bad, the humans could guide it toward answers that were better overall.

But those reviewers weren't always consistent in how they scored the system, and weeding out the problem images from DALL-E 2's training data could also be like a game of whack-a-mole. At first, OpenAI's researchers tried removing all the overly sexualized images of women they could find in the training set so that Dall-E 2 wouldn't portray women as sexual objects. But doing that had a price: it cut the number of women in the dataset "by quite a lot," according to OpenAI's head of research and product at the time, Mira Murati. She doesn't say by how much. "We had to make adjustments because we don't want to lobotomize the model. It's really a tricky thing."

DALL-E 2's photorealistic faces were its biggest liability when it came to stereotypes, and OpenAI seemed fully aware of the problem. So much so that when an internal group of four hundred people started testing the system—mostly OpenAI and Microsoft employees—OpenAI banned them from publicly sharing any of DALL-E 2's realistic portraits.

Some of OpenAI's employees worried about the speed at which OpenAI was releasing a tool that could generate fake photos. Having started off as a nonprofit devoted to safe AI, it was turning into one of the most aggressive AI companies on the market. One anonymous member of the company's team who worked on safety testing told *Wired* that it seemed like the company was releasing the technology to show it off to the world, even though "there's so much room for harm right now."

But Altman's eye was on the bigger prize. He believed the new system had crossed an important threshold on the path to AGI. "It seems to really understand concepts," he said in one interview, "which feels like intelligence." DALL-E 2 was so magical that it could make skeptics of AGI start taking the idea seriously, he added.

The magic here wasn't DALL-E 2's capabilities alone. It was the impact the tool was having on people. "Images have an emotional power," he said. DALL-E 2 was generating buzz. And unlike GitHub Copilot, which could finish writing code that someone had already started, this was creating content fully formed, from start to finish. It was like asking a graphic artist for a picture of anything you wanted.

This idea of generating fully formed content was what made Altman's next move even more sensational. GPT-1 had been more like an autocomplete tool that continued what a human started typing. But GPT-3 and its latest upgrade, GPT-3.5, created brand-new prose, just like how DALL-E 2 made images from scratch.

As the world gawked at DALL-E 2, rumors swirled that rival Anthropic was working on a chatbot, sparking the competitive juices at OpenAI. In early November 2022, OpenAI managers told staff that they were going to launch a chatbot of their own in just a few weeks, that was built on GPT-3.5. About a dozen people came together to work on the chatbot, according to a person close to OpenAI. It wasn't all that different from Google's Meena, which Noam Shazeer had worked on two years earlier, but which Google had kept under wraps.

This wouldn't be a product launch, OpenAI's leadership assured staff, but a "low-key research preview." Still, some employees said they weren't comfortable releasing the tool so quickly. They didn't know how the public might misuse a language model that was so fluent and capable.

Not only that, the chatbot often made factual errors. The researchers working on it decided not to make the system more cautious because that caused it to decline questions it could answer correctly. They didn't want it to say, "I don't know." Instead, they calibrated it to sound more authoritative, even though this meant the chatbot would spout mistruths at least some of the time. They named it ChatGPT.

Altman pushed to launch. He argued that hundreds of OpenAI staff had already tested and vetted ChatGPT, and that it was important to acclimate humanity to what artificial intelligence was destined to do, like dipping your toes into a cold swimming pool. In a way, OpenAI was doing the world a favor and getting it ready for OpenAI's more powerful, upcoming model, GPT-4. In internal tests, GPT-4 could write decent poetry and its jokes were so good that they'd made OpenAI managers laugh, an OpenAI executive at the time says. But they had no idea what kind of impact it would have on the world or society, and the only way to know was to put it out there. On its website, OpenAI called this its "iterative deployment" philosophy, releasing products into the wild to better study their safety and impact. It was the best way to ensure it was building AGI for the benefit of humanity, the company said.

On November 30, 2022, OpenAI published a blog post announcing a public demo of ChatGPT. Many people at OpenAI, including some who worked on safety, weren't even aware of the launch, and some started taking bets on how many people would use it after a week. The highest estimate was one hundred thousand users. The tool itself was just a website with a text box. You typed anything you wanted into the box, and the bot behind it all would respond. It was powered by GPT-3.5. Most of the public hadn't heard of OpenAI, never mind GPT-3. And no one, including researchers at OpenAI, knew what would happen when they let anyone test its capabilities.

"Today we launched ChatGPT," Altman tweeted at about 11:30 a.m. San Francisco time. "Try talking with it here: http://chat.openai.com."

At first, there was silence as a niche audience of software developers and scientists jumped onto the site and started trying it out. Within the next few hours, their reviews started popping up on Twitter:

> 12.26 PT @MarkovMagnifico: playing with ChatGPT [right now] and I've now moved my AGI timeline up to today

> 12:37 PT @AndrewHartAR: ChatGPT just got released. I've seen the future.

> 13:37 PT @skirano: Absolutely insane. I asked #chatGPT to generate a simple personal website. It showed step by step … how to create it, then added HTML and CSS.

> 14:09 PT @justindross: ChatGPT is immediately a better starting point for me than Google for questions I have. This is pretty crazy.
>
> 14:29 PT @Afinetheorem: you can't give take-home essays/assignments anymore.

It was hard to find a single negative appraisal of ChatGPT. The overwhelming response was awe. What made it all the more remarkable was not just how fluent it was but how broad its knowledge seemed to be. Most everyone had tried a chatbot before, whether it was Alexa or a customer service bot of some sort, and most were used to limited, stumbling chats. But ChatGPT could answer pretty much anything eloquently. It was like going from talking to a toddler to a full-blown adult with a college degree.

Within the next twenty-four hours, more and more people piled onto ChatGPT, straining its servers and testing its limits. Now it was everyday professionals, tech workers, people in marketing and the media, who were road testing the bot. Itching to make a spectacle on Twitter, they turned their experiments into a public competition to see who could get ChatGPT to write the funniest, smartest, or weirdest text. It was like the fanfare around DALL-E 2 all over again, but bigger. Over the next few days, people flooded Twitter with screenshots of ChatGPT's poems, raps, sitcom scenes, and emails. The more offbeat, the better.

One Twitter user named Thomas H. Ptacek asked it to "write a biblical verse in the style of the King James Bible, explaining how to remove a peanut butter sandwich from a VCR."

OpenAI's bot complied with the following:

> *And it came to pass that a man was troubled by a peanut butter sandwich, for it had been placed within his VCR, and he knew not how to remove it.*
>
> *And he cried out to the Lord, saying, "Oh, Lord, how can I remove this sandwich from my VCR, for it is stuck fast and will not budge?"*

"I'm sorry, I simply cannot be cynical about a technology that can accomplish this," Ptacek tweeted. Within a week, more than a million people had used ChatGPT. After two months, ChatGPT had attracted thirty million registered users, making it one of the fastest-growing online services in history. By early 2024, around one hundred million people were using ChatGPT weekly. No standalone AI tool had ever reached that kind of mainstream popularity before.

On March 14, 2023, the very same day that Anthropic had finally released its own chatbot called Claude, OpenAI launched its upgrade, GPT-4. Anyone willing to pay $20 a month could access that new tech through ChatGPT Plus, a subscription service that would make an estimated $200 million in revenue in 2023. Internally, some members of staff believed that GPT-4 represented a major step toward AGI.

Machines weren't just learning statistical correlation in text, Sutskever said in one interview. "This text is actually a projection of the world.… What the neural network is learning is more and more aspects of the world, of people, of the human condition, their hopes, dreams and motivations, their interactions and the situations that we are in."

"Once you have a system that can take in observations about the world, learn to make sense of them—and one way to do that is to predict what's going to happen next—I think that is very near intelligence," Altman said in another interview.

The tech press were captivated. The *New York Times* called ChatGPT "the best artificial intelligence chatbot ever released to the general public." Journalists who tried the system found themselves charmed by the system's friendly and enthusiastic responses. On Twitter, some tech enthusiasts boasted about how they were already using it to draft their emails or other work-related documents to make themselves more productive.

Naturally, that sparked a new wave of press articles about whether ChatGPT would replace humans. Altman went on a publicity tear to address all the excitement and meet people's concerns head-on via podcasts, newspapers, and other news publications. Yes, he said, this was probably going to replace jobs—think copy writers, customer service operators, and even software developers—but that didn't mean ChatGPT and the technology underpinning it would replace human work altogether.

"Some jobs are going to go away," Altman said bluntly in one interview. "There will be new, better jobs that are difficult to imagine today." This was met with a quiet resignation among the press and general public, because historic shifts like the Industrial Revolution had shown that technology could indeed bring painful changes to employment. And generative AI systems like ChatGPT weren't flash-in-the-pan fads like crypto. ChatGPT was useful. People were already ginning up high school essays, brainstorming business plans, and conducting marketing research with it.

Inside OpenAI, staff consoled themselves that the future would be worth it, arguing that the transition to machine-operated work and factories during the Industrial Revolution had also led to new jobs and better standards of living. But a divide was also growing between OpenAI employees who were focused on product development and those focused on safety, who were struggling to monitor the soaring incoming traffic on ChatGPT for abusive queries. Believing they were taking significant steps toward AGI, Ilya Sutskever began working more closely with the company's safety team. Even so, OpenAI's product team doubled down on commercializing ChatGPT, inviting businesses to pay for access to its underlying technology.

Inside Google, executives recognized that more and more people might just go to ChatGPT for information about health issues or product advice—among the most lucrative search engine terms to sell ads against—instead of Google.

Google arguably deserved some proper competition. Over the years its results page had become cluttered with ads and sponsored links as it tried to squeeze as much revenue out of each individual search as it could, even if that made its product more unpleasant to use. If it could confuse people about what was an ad versus what was an actual search result, it could make more money.

Between 2000 and 2005, Google had marked ads more clearly, giving them a blue background and ensuring they only took up one or two links at the top of the page. But over the years, it became harder to tell the difference between ads and normal web links. The blue background faded to green, then to yellow, and then to nothing it all. Ads started taking up more of the page, forcing people to scroll for longer to find those proper results. As annoying as this was for consumers, Google could get away with it because internet users didn't think they had anywhere else to go. More than 90 percent of online searches around the world happened on Google.

But now, for the first time, Google's more-than-twenty-year dominance as gatekeeper to the web was on shaky ground. For years, its main moneymaker had been a system that crawled billions of web pages and indexed and ranked them to find the most relevant answers to queries. It then churned out a list of links to click through. But ChatGPT offered something more tantalizing for busy internet users: a single answer that was based on its own synthesis of all that information. No endless scrolling or searching through a maze of ads and links. ChatGPT did all that for you.

Take, for instance, a query about whether condensed milk or evaporated milk was better for pumpkin pie. If you asked ChatGPT, you'd get a single detailed answer about how condensed milk was probably superior because it would lead to a sweeter pie. Google would spit out a long list of links to ads, recipes, and articles you'd have to click around and read. The infinite possibilities that had once made Google so remarkable were now just a time suck. In Silicon Valley, technologists were forever chasing the "frictionless" online experience. A frictionless alternative to Google posed a potential financial disaster to the company.

Within weeks of ChatGPT's launch, executives at Google issued a code red inside the company. The company had been caught on its heels and badly. Since 2016, Chief Executive Sundar Pichai had been calling Google "AI-first." So how had a little company with fewer than two hundred AI researchers developed something better than what Google had with nearly five thousand? The threat was made more serious by OpenAI's close ties with the deep-pocketed Microsoft.

Google already had LaMDA, the older language model that its engineer had thought was sentient. But its executives were in a predicament. What if they released a competitor to ChatGPT and people started using that instead of Google search? That meant they wouldn't click around on the ads, sponsored links, and other websites that used Google's ad network and drove its profits.

More than 80 percent of Alphabet's $258 billion in 2021 revenue had come from advertising, with much of that coming from pay-per-click ads that people reached by using its search engine. All those ads that were clogging up Google's search results had become critical to its business. It couldn't just change the status quo. "The goal of Google search is to get you to click on links, ideally ads," says Sridhar Ramaswamy, who ran Google's ads and commerce business between 2013 and 2018. "All other text on the page is just filler."

Google had for years been taking a cautious, almost fearful approach to new technology. It "didn't move" unless something was a billion-dollar business, and it certainly didn't want to mess with its own ads business that made nearly $260 billion a year.

"It gets harder as you get bigger," Ramaswamy says. "At Google, the size of the ad team was typically four to five times the size of the organic search team. To start a product that is the antithesis of the core model is really hard to get done in reality."

But now Google executives didn't have much choice. In one meeting, which was recorded and shared with the *New York Times*, a manager pointed out that smaller companies like OpenAI seemed to have fewer concerns about releasing radical new AI tools to the public. Google had to jump in and do the same, or it risked becoming a dinosaur. Putting caution aside, everything went into high gear.

Panicked executives told staff working on key products that had at least one billion users, like YouTube and Gmail, that they had just months to incorporate some form of generative AI. Google had been the world's indexing machine for years, processing videos, images, and data, but now it had to start *creating* new data, too, with AI. Making this kind of fundamental shift was like trying to drive a spluttering old truck that only ever went twenty miles an hour onto a race car track. Executives were so desperate that they summoned Google founders Larry Page and Sergey Brin, who had resigned as co-CEOs of Alphabet back in 2019, to help figure out a response to ChatGPT in a series of emergency meetings.

Sensing deep insecurity from Google's leadership, the company's engineering teams delivered. A few months after the launch of ChatGPT, managers at YouTube added a feature where video creators on the website could generate new film settings or swap outfits, using generative AI. But it felt like they were throwing spaghetti at the wall. It was time to bring out their secret weapon: LaMDA.

Pichai sent a company-wide memo telling his employees to test out a new chatbot that they would soon release to the public and rewrite any answers that they deemed bad. He then published a blog post on February 6, 2023, telling the world that something new was on its way. Under the title "An Important Next Step in Our AI Journey," he wrote, "We've been working on an experimental conversational AI service, powered by LaMDA, that we're calling Bard."

Eager to stay in the lead, Microsoft posted an announcement the following day. Bing, its backwater search engine that had a piddling 6 percent share of the market for online queries, would soon get a big AI upgrade. OpenAI's latest GPT language model would power Bing to "unlock the joy of discovery, feel the wonder of creation, and better harness the world's knowledge." Translation: it could do what ChatGPT was already doing but with certain advancements that only Microsoft knew about.

This breathless race to launch was wowing the world, until a few close watchers noticed some glitches. Google had posted examples of clever answers from Bard and Microsoft from Bing. But when a few journalists double-checked some of those answers, it turned out they were wrong. In a launch video shown by Pichai, Bard botched a historical fact about the James Webb telescope, while Bing misstated some earnings numbers from retailer The Gap.

The chatbots weren't just hallucinating facts but suffering from some kind of mood disorder too. Not long after Microsoft's announcement, *New York Times* writer Kevin Roose published a column about an unsettling two-hour conversation he'd held with Bing late one night, where Microsoft's new search engine turned chatbot confessed its love for the writer and insisted that "you're not happily married." Roose wrote that the encounter had given him a "foreboding feeling that A.I. had crossed a threshold, and that the world would never be the same."

For Microsoft's Nadella, all this hype and attention to Bing translated into a delicious opportunity to gloat. He told one interviewer that he'd been waiting for years for the chance to challenge Google's dominance of search and that now Bing could finally pull it off. "And I want people to know that we made them dance," he added.

From the outset, none of this made sense. Google had done everything early. Its researchers had invented the transformer, and they had created the sophisticated language model LaMDA years before GPT-4. Its own AI lab, DeepMind, had set off on a mission to build AGI five years *before* OpenAI had even been founded to do the same. Yet Google was now racing to catch up.

Its lumbering bureaucracy and fear of disrupting its business and reputation had caused a deep-set inertia. Paradoxically, that had protected the world from some of the risks that OpenAI had now introduced, risks that were most likely to impact minority groups and put a cleaver to large swathes of jobs.

OpenAI's big splash also called into question DeepMind's work over the past thirteen years. And it rattled Hassabis. Weeks after ChatGPT's release, he told staff in an all-hands meeting that DeepMind shouldn't become the "Bell Labs of AI," a place that invented everything but saw its ideas commercialized by others, a former employee remembers.

Meanwhile, no one was asking where AGI was. But they *were* asking where the useful, humanlike AI was. DeepMind had managed to create AI systems that could beat human champions at Go and other games, but OpenAI's ability to create a system that could simply write an email was somehow more impressive.

The scientific strategy that Demis Hassabis had been chasing was starting to look insular. Hassabis had sought to build AGI through games and simulations and measured the success of his company's work through awards and the prestige of publishing papers in scientific journals. OpenAI's approach to AI had been driven by engineering principles and scaling existing technology as much as possible. DeepMind's had been more academic, publishing research papers about the *AlphaGo* gaming system and AlphaFold, a novel approach to predicting how proteins fold in the human body.

AlphaFold was born out of a hackathon—or a collaborative programming event—at DeepMind in 2016, before turning into one of the company's most promising projects.

Hassabis had dreamed of using AGI to solve big global problems like cancer, and it seemed like he finally had an AI system that could do something like that.

When amino acids in our cells fold up into specific 3D shapes, they become proteins, and badly folded proteins can lead to diseases. AlphaFold was an AI program that predicted what those 3D shapes would look like when they folded up, and DeepMind believed that could help scientists better understand what kinds of chemical reactions might affect those proteins, aiding drug discovery.

Hassabis made it an urgent priority for DeepMind to win a global competition for protein folding called CASP in 2019 and 2020. “We need to double down and go as fast as possible from here,” he told his staff in one meeting that was captured in a video documentary. “We’ve got no time to lose.”

While Altman measured success with numbers, whether for investments or people using a product, Hassabis chased awards. He often told staff that he wanted DeepMind to win between three and five Nobel Prizes over the next decade, according to people who worked with him.

DeepMind won at CASP in both 2019 and 2020 and open-sourced its protein folding code to scientists in 2021. At the time of writing, more than one million researchers across the world had accessed the AlphaFold Protein Structure Database, according to DeepMind. But science is a slow process, and while Hassabis could one day still win a Nobel, a major discovery using his system remains elusive. Some experts are also skeptical that DeepMind’s protein shape predictions are accurate enough to reliably identify how drug compounds will bind to proteins or that it could save that much time in drug discovery.

All told, DeepMind’s biggest projects had garnered lots of prestige but made relatively little impact on the real world. It had insisted on training AI in fully simulated environments, where physics and other details could be precisely designed and fully observed. That was how it built *AlphaGo*, by programming it to play millions of games against itself in simulation, and AlphaFold, which used simulations of protein folding.

Training on real-world data—as OpenAI had done by scraping billions of words from the internet—was messy and noisy. It left them open to scandal, as Hassabis had learned through the hospitals project. But DeepMind’s self-contained approach also meant it was harder to build AI systems that people could use in the real world.

Hassabis had become so focused on the virtual worlds of his AI systems and on chasing recognition that he missed the revolution in language models. Now he had to follow in Altman’s footsteps. Google executives told DeepMind to start working on a series of large language models that would be even better than LaMDA. They called the new system Gemini, and DeepMind imbued it with the strategic planning techniques that *AlphaGo* had developed.

To help move things along more quickly, Pichai made another drastic move. He merged the two rival AI divisions, DeepMind and Google Brain, and called them Google DeepMind. (Staff called the new unit GDM for short.) Having spent years competing with each other to hire top researchers and fighting for more computing power, the two units also had completely different cultures. While Google Brain was closer to the mothership and worked directly on improving Google products, DeepMind was independent to the point of being aloof—its staff wore badges that gave them access to other Google buildings, but Google staff couldn't get into DeepMind's, for instance.

To the surprise of many, Pichai picked Hassabis to run the combined unit. Jeff Dean, Google's most revered engineer who oversaw AI research at the rest of the company, had seemed like the more likely candidate. Instead, the former game designer and simulation obsessive, the guy who had spent years trying to split away, was now leading Google's big project to protect its lead in web search. Politically he was wielding more power than ever before, and by controlling more of Google, he could control more of DeepMind again too.

"Demis's profile and influence in Google is much more now than it was a few years ago," says Shane Legg. "Instead of becoming a bit more independent, we became integral to Google itself. It's critical for us and our mission that Google is successful.

"That wasn't obvious to me a few years ago," he adds. "I thought we might need a bit more independence. In hindsight, I think what actually happened may be better."

When Hassabis announced the merger with Google Brain to DeepMind staffers, he told them in an email that the units were joining forces because AGI had the potential to "drive one of the greatest social, economic and scientific transformations in history."

In reality, they were merging to help a panicked Google beat a business rival, just as OpenAI's mission to benefit humanity (without "financial pressure") had shifted toward serving the interests of Microsoft. The so-called mission drift that was so common in Silicon Valley, as it had been with WhatsApp, was happening to technology that could have far greater influence on society. OpenAI tried to address that in July 2023, when it announced that Ilya Sutskever would lead its new Superalignment Team. Within four years, the company said, Sutskever's researchers would figure out how to control AI systems as they became smarter than humans.

But OpenAI still had a glaring problem. It was sidestepping the need for transparency, and more broadly, it was getting harder to hear the voices calling for more scrutiny of large language models. Gebru, Mitchell, and Bender, whose notorious research paper had finally drawn attention to the risks, were still trying to warn the public about how those models, and generative AI more generally, could perpetuate stereotypes. Unfortunately, governments and policymakers were paying more attention to a well-financed group of louder voices: the AI doomers.

CHAPTER 14

A Vague Sense of Doom

Sam Altman had set off several different races when he launched ChatGPT. The first was obvious: Who would bring the best large language model to market first? The other was taking place in the background: Who would control the narrative about AI?

In March 2023, a few weeks after Microsoft and Google made their hasty launches of Bing and Bard, Eliezer Yudkowsky wrote a two-thousand-word column in *Time* magazine about where AI was headed, painting a terrifying picture of a future with more intelligent machines.

“Many researchers steeped in these issues, including myself, expect that the most likely result of building a superhumanly smart AI, under anything remotely like the current circumstances, is that literally everyone on Earth will die,” he wrote.

That same month, an open letter signed by Elon Musk and other technology leaders called for a six-month “pause” on AI research because of the risks to humanity. “Should we develop nonhuman minds that might eventually outnumber, outsmart, obsolete and replace us?” said the letter, which was put together by Jaan Tallinn’s Future of Life Institute. “Should we risk loss of control of our civilization?” The letter, which had nearly thirty-four thousand signatories, grabbed headlines around the world from news outlets including Reuters, Bloomberg, the *New York Times,* and the *Wall Street Journal*.

Further breathless coverage was given to two AI researchers, deemed “godfathers” of AI—Geoffrey Hinton and Yoshua Bengio—after they warned the press about AI’s existential threat to the human race. Bengio said he felt “lost” over his life’s work, and Hinton said he regretted some of his research.

“The idea that this stuff could actually get smarter than people—a few people believed that,” he told the *New York Times*. “But most people thought it was way off. And I thought it was way off. I thought it was 30 to 50 years or even longer away. Obviously, I no longer think that.… I don’t think they should scale this up more until they have understood whether they can control it.”

AI’s biggest names all seemed to be saying the same thing: AI development was moving too fast and could spin out of control in a catastrophic way. The idea of an extinction threat from AI was becoming a fixture in public discourse, so much so that you could bring it up with your in-laws at dinner and they’d be nodding along at its importance. The mainstream public found themselves entranced by the idea that we could have machine overlords that went rogue. By late 2023, about 22 percent of

Americans believed that AI would cause human extinction in the next fifty years, according to a poll of about 2,444 US adults by market research firm Rethink Priorities.

Yet all this talk of doom had a paradoxical effect on the business of AI itself: it was booming. Funding for start-ups that built generative AI products soared in 2023 to more than $21 billion, from about $5 billion a year earlier, according to Pitchbook, a market research firm.

The implicit message of rogue AI was enticing. If this technology might destroy the human race in the future, didn't that also mean it was powerful enough to boost your business now?

And it seemed like the more Sam Altman talked about the threat of OpenAI's technology—telling Congress, for instance, that tools like ChatGPT could "cause significant harm to the world"—the more money and attention he attracted. In January 2023, OpenAI secured another investment from Microsoft, this time worth $10 billion, in exchange for granting the software giant a 49 percent stake in the firm. Microsoft was now as close as you could get to controlling OpenAI outright.

Anthropic, the new company that Dario Amodei and a group of other researchers from OpenAI had funded, were also attracting big investments. By late 2023, it had accepted a $2 billion investment from Google and a $1.3 billion investment from Amazon. Within a year, its value had quadrupled to more than $20 billion. It seemed that making super AI that was supersafe could also make you supervaluable. Behind the scenes, Anthropic wanted to raise as much as $5 billion to enter more than a dozen industries and challenge OpenAI, according to company documents obtained by TechCrunch. "These models could begin to automate large portions of the economy," Anthropic's documents said, adding that this was a race in which Anthropic could stay ahead for many years if it could build "the best" models by 2026.

Safety-first framing had made Anthropic sound like a nonprofit, with its mission to "ensure transformative AI helps people and society flourish." But OpenAI's smash hit with ChatGPT had shown the world that the companies with the grandest plans could also be the most lucrative investments. Proclaiming that you were building safer AI had almost become like a dog whistle for bigger tech companies who wanted to get in on the game too.

Anthropic would twist itself in knots to explain this logic. In order to figure out how to make AI systems safer, it couldn't just study the world's most powerful AI systems—it had to build them. Hence the wink and nod to large technology companies who were Earth's sole proprietors of massive computing power. As part of Anthropic's deal with Google, for instance, it would get cloud computing credits that would let it build a large language model that would rival OpenAI's.

In public there were now two different groups of people calling for safer AI. There were those like Altman and Amodei who had signed yet another open letter stating that "mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war." They came under the umbrella of “AI safety,” painting the future threat in vague terms and rarely spelling out what rogue AI systems would do or when it would happen. They also tended to advocate for light-touch regulation when they brought those concerns before Congress.

The other group included those like Timnit Gebru and Margaret Mitchell, who’d been agitating for years over the risks that AI already posed to society. This “AI ethics” group tended to skew toward women and people of color who had firsthand experience of stereotyping and who feared that AI systems would continue to perpetuate inequality. Over time, they became increasingly outraged by the actions of those in the “AI safety” camp, not least because that group was making so much money.

The funding disparity was stark. The ethics side was often scrambling for cash. Groups like the European Digital Rights Initiative, a twenty-one-year-old network of nonprofit groups that campaigned against facial recognition and biased algorithms, had an annual budget of just $2.2 million in 2023. Similarly, the AI Now Institute in New York, which scrutinized how AI was used in healthcare and the criminal justice system, had a budget of less than $1 million.

Groups that were focused on AI “safety” and the extinction threat got far more funding, often via billionaire benefactors. The Future of Life Institute, a Cambridge, Massachusetts–based nonprofit that studied how best to stop AI from getting access to weapons, got $25 million from crypto magnate Vitalik Buterin in 2021. That single grant was bigger than the combined annual budgets of all the AI ethics groups at the time.

Open Philanthropy, the charitable vehicle of Facebook billionaire Dustin Moskovitz, has sprinkled a number of multimillion-dollar grants to AI safety work over the years, including a $5 million donation to the Center for AI Safety in 2022 and an $11 million donation to Berkeley’s Center for Human-Compatible AI.

All told, Moskovitz’s charity has been the biggest donor to AI safety, by virtue of the near $14 billion fortune that he and his wife, Cari Tuna, plan to mostly give away. That includes a $30 million donation to OpenAI when it first established itself as a nonprofit.

Why has so much money gone to engineers tinkering on larger AI systems on the pretext of making them safer in the future, and so little to researchers trying to scrutinize them today? The answer partly comes down to the way Silicon Valley became fixated on the most efficient way to do good and the ideas spread by a small group of philosophers at Oxford University, England.

Back in the 1980s, an Oxford philosopher, Derek Parfit, started writing about a new kind of utilitarian ethics, one that looked far into the future. Imagine, he said, that you left a broken bottle on the ground and one hundred years later, a child cuts their foot on it. They might not yet be born, but you would shoulder the same burden of guilt as if that child was injured today.

“His very simple basic thought is that morally, future people are as important as present people,” says David Edmonds, who wrote a 2023 biography about Parfit. “Imagine these three scenarios. A, there’s peace. B, 7.5 billion of the 8 billion people in the world are exterminated in a war. And C, everyone is killed. Most people’s intuition is that the gap between A and B is much bigger than the gap between B and C. But Parfit says that's wrong. The gap between B and C is much more significant than the gap between A and B. If you wipe out the whole of humanity, you wipe out all future generations."

Here's one way to quantify that. Mammals have an average species "life span" of about one million years, and humans have been around for about two hundred thousand years. That theoretically gives us another eight hundred thousand years on the planet. If the current world population stabilizes at eleven billion people, based on United Nations projections for the end of this century, and the average life span rises to eighty-eight years, that could mean, according to one estimate, that another *one hundred trillion people* have yet to live in the future.

To help visualize those numbers, imagine a small dinner knife and a solitary pea are sitting on your dining room table. The knife represents the number of people who have already lived and died in the past. The pea is everyone who is alive today. The surface of the dining table is the number of people who have yet to live—and it could be much bigger, if humans prove themselves to be longer-lasting than your typical mammalian species.

In 2009, an Australian philosopher named Peter Singer expanded on Parfit's work with a book called *The Life You Can Save*. Here now was a solution: wealthy people should not just donate money based on what felt right but use a more rational approach to maximize the impact of their charitable giving and help as many people as possible. By helping many of those yet-to-be-born people in the future, you could be even more virtuous.

These ideas started to make the leap from academic papers to the real world and form the basis of an ideology in 2011, when a twenty-four-year-old Oxford philosopher named Will MacAskill cofounded a group called 80,000 Hours. The number referred to the average hours a person works in their lifetime, and the organization targeted college campuses in the United States, advising young university graduates on careers that would have the greatest moral impact. It often steered the technically minded ones toward AI safety work. But the group also encouraged graduates to pick careers that paid the highest salaries, allowing them to donate as much money as possible to high-impact causes.

MacAskill and his young team eventually reincorporated themselves as the Center for Effective Altruism and a new credo was born. The driving idea behind effective altruism was efficiency. People who lived in wealthy countries had an obligation to help those in poorer nations because that's where they could have the most bang for their buck. You could help more people in Africa through global health charities, for instance, than by donating to the poor in America. It was also morally better to spend your time earning as much money as possible so that you could be like Dustin Moskovitz and give lots of it away. When he gave talks to students, MacAskill would show a slide asking if they could do more good as a doctor or a banker. His answer was that it was better to become a banker. You might be able to save a certain number of lives in Africa as a doctor, but as a banker you could hire *several* doctors to save many more lives.

This offered graduates a counterintuitive way of looking at all the inequalities of modern capitalism. Now there was nothing wrong with a system that allowed a handful of humans to become billionaires. By amassing unfathomable amounts of wealth, they could help more people!

The movement picked up its biggest name in 2012, when MacAskill reached out to someone whom he hoped to recruit to the cause, an MIT student with dark curly hair named Sam Bankman-Fried. The two had coffee, and it turned out that Bankman-Fried was already a fan of Peter Singer and interested in causes related to animal welfare.

MacAskill steered Bankman-Fried away from the idea of working directly with animal causes and said he could help them much more by going into a high-earning field. Bankman-Fried was immediately hooked, according to Michael Lewis's account of his rise and fall in the book *Going Infinite*. "What he said sort of seemed obviously right to me," Bankman-Fried said in the book. He took a job at a quantitative trading firm and eventually founded the crypto-currency exchange FTX in 2019.

Bankman-Fried put effective altruism front and center of that business. His cofounders and management team were effective altruists and kept MacAskill on as a member of the FTX Future Fund, which would go on to give $160 million to effective altruism causes in 2022, some of which were directly linked to MacAskill. He frequently talked to the press about giving all his money away, and in large posters advertising FTX, he was pictured in his trademark T-shirt and cargo shorts, flanked by the words: "I'm in on crypto because I want to make the biggest global impact for good." He positioned himself as an ascetic character who, despite his billionaire status, drove a Toyota Corolla, lived with roommates, and often looked disheveled.

Many technologists saw this approach to morality as a breath of fresh air. When engineers saw a problem, they often solved it formulaically, debugging code and optimizing software through constant testing and evaluation. Now you could also quantify moral dilemmas, almost like they were math. People in effective altruism circles sometimes talked about maximizing the effect of a charitable act by focusing on "expected value," a number you got by multiplying the value of an outcome by the probability that it would occur.

As effective altruism took greater hold in Silicon Valley, its focus shifted from buying cheap malaria nets and helping as many people as possible in Africa, to issues with a more science fiction flavor. Elon Musk, who tweeted that MacAskill's 2022 book was a "close match for my philosophy," had wanted to send people to Mars to ensure the long-term survival of humans. And as artificial intelligence systems became more sophisticated, it made sense to keep it from going rogue and wiping out humanity too. Many of the staff at OpenAI, Anthropic, and DeepMind were effective altruists.

Acting on the extinction risk of AI is a rational calculation. Even if there is only a 0.00001 percent risk that AI might extinguish humanity, that cost is so big it is essentially infinite. If you multiply tiny odds with an infinite cost, you still get a problem that is infinitely large. This rationale is all the more potent if you believe, as some AI safety advocates do, that computers of the future will host the conscious minds of billions of people and also create new forms of sentient, digital lives. Those one hundred trillion people who have yet to live in the future could be a much higher number. Following this kind of moral math to the letter, it makes sense to prioritize the tiny possibility of having to save more than one hundred trillion physical and digital lives from destruction. Global poverty is a rounding error by comparison.

After OpenAI launched in 2015, funding poured into AI extinction causes. Moskovitz's Open Philanthropy increased the number of grants it was giving to issues relating to so-called long-termist causes including AI safety research, from $2 million in 2015 to more than $100 million in 2021.

Bankman-Fried had jumped in too. His FTX Future Fund, run by effective altruists like Nick Beckstead and MacAskill, pledged to donate $1 billion to projects aiming to "improve humanity's long-term prospects." When it listed the fund's areas of interest, it started with "the safe development of artificial intelligence."

When *New Yorker* magazine profiled life inside the Future Fund, it noted that office chitchat at its Berkeley, California, headquarters often veered toward when an AI apocalypse might happen.

"What are your timelines?" staff would ask one another. "What's your p(doom)?"

P stood for probability and the question referred to how people quantified the risk of an AI doomsday. Someone with a more optimistic outlook might put their p(doom) at 5 percent. Ajeya Cotra, a research analyst at Open Philanthropy who helped decide grant-making, told one podcast that hers was between 20 and 30 percent.

Nobody knew Bankman-Fried's p(doom), but he cared enough about AI safety to invest $500 million in Anthropic. His FTX cofounders and fellow effective altruists, Nishad Singh and Caroline Ellison, also invested in the start-up that had split from OpenAI about a year earlier.

In early 2022, MacAskill noticed a tweet from Musk, saying that he wanted to buy Twitter to save free speech. The Scottish philosopher sent Musk a text. At the time, Bankman-Fried was worth $24 billion, making him one of the richest effective altruists on earth. But Musk's $220 billion fortune could singlehandedly make effective altruism the world's biggest philanthropic movement.

MacAskill told Musk that Bankman-Fried also wanted to buy Twitter to help make it "better for the world." Did the two want to combine their efforts?

"Does he have huge amounts of money?" Musk texted back.

"Depends on how you define 'huge'!" MacAskill replied, according to court documents. MacAskill said Bankman-Fried could contribute as much as $8 billion.

"That's a start," Musk replied.

"Would you like me to introduce you two via text?" MacAskill asked.

Musk didn't answer the question. "You vouch for him?" he asked.

"Very much so!" MacAskill replied. "Very dedicated to making the long-term future of humanity go well."

"Ok then sure."

"Great!"

Although Musk eventually connected with Bankman-Fried, they never came to a financial agreement, which meant Musk dodged a bullet. Months later, FTX collapsed amid rumors that Bankman-Fried had been fraudulently transferring client funds inside the company. At trial, prosecutors accused him of swindling $8 billion from thousands of customers and investors, and he faced decades in prison. Having framed himself as an ascetic, it turned out Bankman-Fried had been living in a luxurious penthouse in the Bahamas while throwing hundreds of millions of dollars at various investments. Now much of the money he'd earmarked for effective altruism had gone up in smoke, and it also transpired that he hadn't been that enthusiastic about it anyway.

Soon after FTX's collapse, Bankman-Fried gave a remarkable interview with news site Vox:

"So the ethics stuff—mostly a front?" the reporter asked.

"Yeah," Bankman-Fried replied.

"You were really good at talking about ethics, for someone who kind of saw it all as a game with winners and losers," the reporter noted.

"Ya," said Bankman-Fried. "Hehe. I had to be."

FTX's downfall cast a huge shadow over effective altruism's reputation and became an allegory for some of the movement's fundamental problems. The first was predictable. When people embarked on a mission to do the most good while also seeking the most wealth, they were probably making themselves more susceptible to corrupt behavior and foolhardy, ego-driven judgments. Buying Twitter, for instance, didn't tick any obvious boxes for helping humanity in the long term, but Bankman-Fried was ready to spend as much as $8 billion to buy the site with Musk and stand on a pedestal with the world's richest man, as an act of effective altruism.

After FTX imploded, MacAskill took to Twitter to do damage control: "A clear-thinking [effective altruist] should strongly oppose 'ends justify the means' reasoning," he tweeted. Yet the movement's own principles incentivized people like Bankman-Fried to reach their goals by whatever means necessary, even if that meant exploiting people. It created a myopia that affected even an intelligent Oxford academic like MacAskill, who had chosen to attach himself to someone running a crypto exchange, knowing full well that crypto businesses were speculative at best and a dangerous form of gambling at worst.

Bankman-Fried could rationalize his duplicity because he was working toward a bigger goal of maximizing human happiness. Musk could wave off his own inhumane actions, from baselessly calling people pedophiles on Twitter to alleged widespread racism at his Tesla factories, because he was chasing bigger prizes, like turning Twitter into a free speech utopia and making humans an interplanetary species. And the founders of OpenAI and DeepMind could rationalize their growing support for Big Tech firms in much the same way. So long as they eventually attained AGI, they would be fulfilling a greater good for humanity.

Technologists like Altman and Hassabis knew that the societal problems they hoped to fix with AGI were messy and tangled. That's why so many of them embraced some or all of effective altruism. It offered a simpler, more rational path to solving moral problems while allowing them to make as much money as possible. Billionaires weren't the cause of global poverty but the solution.

It also made it easier to disassociate from humanity. A popular phrase in effective altruism is "shut up and multiply," which means that when making ethical decisions, you should maximize your output by setting aside personal emotions or moral intuitions. For all the devotion to humanity by effective altruists, many like Altman emotionally detached themselves from the world around them, the better to focus on their mission. Within the effective altruist bubble, people worked together, socialized together, funded one another, and had romantic relationships together.

When Open Philanthropy pledged $30 million to OpenAI in 2017, the charity was forced to disclose that it was getting technical advice from Dario Amodei, who was then a senior engineer at the nonprofit. It also admitted that Amodei lived in the same house as Open Philanthropy's executive director Holden Karnofsky. And it further admitted that Karnofsky was engaged to Dario's sister Daniela, who also worked at OpenAI. All of them were effective altruists. It was an incestuous circle.

The movement was insular and increasingly opaque, and so were the AI firms like OpenAI, DeepMind, and Anthropic, staffed by many of its followers. Probably one of the best things these companies could do to stop AI from going rogue was make their AI systems more transparent, as Gebru and Mitchell had pushed for. After all, how would future humans stop AI from going rogue if they lacked the expertise to scrutinize its mechanics, if researchers had been shut out of studying their training data and algorithms for decades? In other words, the transparency that AI ethics campaigners were pushing for today would also address the extinction threat of tomorrow.

OpenAI's argument that it had to remain secretive to stop bad actors from misusing its technology didn't hold much water. It had given itself the all-clear to launch GPT-2 in November 2019 because it saw "no strong evidence of misuse." If that was true, why not release its training data details? More likely because Altman wanted to protect OpenAI from competitors and lawsuits. If OpenAI became more transparent, it would be easier for rivals—not bad actors—to copy their models and reveal the extent to which OpenAI had scraped copyrighted work too.

Altman and Hassabis had started their companies with grand missions to help humanity, but the true benefits they had brought to people were as unclear as the rewards of the internet and social media. More clear were the benefits they were bringing to Microsoft and Google: new, cooler services and a foothold in the growing market for generative AI.

Microsoft had turned Copilot, the AI assistant built on OpenAI's technology, into a wide-ranging service for Windows, Word, Excel, and business-focused software Dynamics 365. Analysts estimate that OpenAI's technology could generate billions in annualized revenue for Microsoft by 2026. At one point in late 2023, when Nadella shared a stage with Altman and was asked about how Microsoft's relationship with OpenAI was going, he burst into uncontrollable laughter. The answer was so obvious it was hilarious. Of course the relationship was going well.

Microsoft was happily splashing more money on its growing AI business and planned to spend more than $50 billion in 2024 and beyond expanding its vast data centers, the engines that powered generative AI. That would make it one of the biggest infrastructure buildouts in history, as Microsoft outspent government projects on railroads, dams, and space programs. Google was expanding its data centers too.

By early 2024, everyone from media to entertainment companies to Tinder were stuffing new generative AI features into their apps and services. The generative AI market was projected to expand at a rate of more than 35 percent annually to hit $52 billion by 2028. Entertainment firms said they could generate content more quickly for films, TV shows, and computer games. Jeffrey Katzenberg, the cofounder of DreamWorks Animation and the producer of *Shrek* and *Kung Fu Panda*, said generative AI would cut the cost of animated movies by 90 percent. "In the good old days, you might need 500 artists and years to make a world-class animated movie," he said at a Bloomberg conference in November 2023. "I don't think it will take 10 percent of that three years from now."

Generative AI would make advertising even more eerily personal. For years, ads could target large groups of people at once; now they could zero in on just one person with hyperpersonalized video ads that could state your name. The World Economic Forum said that large language models would enhance jobs that required critical thinking and creativity. Anyone from engineers to ad copywriters to scientists could use them as extensions of their brains. And governments were upgrading their AI systems to assess welfare claims, monitor public spaces, or determine someone's likelihood of committing a crime.

Google, Microsoft, and a new generation of start-ups were racing to capture as much of that new business as they could, seeking an edge over their competitors. Close to half of American corporate board members called generative AI the "main priority above anything else" for their companies, according to a late 2023 survey by *Fast Company*. Here, for instance, was how the CEO of Bumble described the dating app's main plans for 2024: "We really want to embark big on AI," she said. "AI and generative AI can play such a big role in accelerating people finding the right person."

Bumble wanted to use the tech behind ChatGPT to build personal matchmakers. Instead of ticking a bunch of boxes on the app, you would simply tell its bot everything you wanted in a partner—from your desire to have children, to your political views, to what you did on a typical Saturday morning. The AI matchmaker would then “talk” to the AI matchmakers of other Bumble users to find the most compatible human. Instead of swiping through hundreds of different people, AI would do that for you.

As these and other business ideas gathered pace, the price of stuffing generative AI into everything was still unclear. Algorithms were already steering more and more decisions in our lives, from what we read online to who companies wanted to recruit. Now they were poised to handle more of our thinking tasks, which raised uncomfortable questions not only about human agency but also about our ability to solve problems and simply imagine.

Evidence suggests that computers have already offloaded some of our cognitive skills in areas like short-term memory. In 1955, a Harvard professor named George Millar tested the memory limits of humans by giving his subjects a random list of colors, tastes, and numbers. When he asked them to repeat as many things on the list as they could, he noticed that they were all getting stuck somewhere in the neighborhood of seven. His paper, “The Magical Number Seven, Plus or Minus Two,” went on to influence how engineers designed software and how telephone companies broke down phone numbers into segments to help us recall them. But according to more recent estimates, that magic number has now fallen from seven to four.

Some call this the Google Effect. By relying more and more on the search giant to recall facts or give us driving directions, we’ve outsourced our memory to the company and inadvertently weakened our short-term memory skills. Could something like that happen to deeper aspects of our cognition as we become overreliant on AI to generate ideas, text, or art? On Twitter, some software developers have admitted to using it so much to write code that their productivity drops whenever a service like Copilot temporarily goes offline.

History shows humans do tend to fret that new innovations will cause our brains to shrivel up. When writing first became widespread more than two thousand years ago, philosophers like Socrates worried it would weaken human memory because before its advent, it was only possible to pass on knowledge through spoken discourse. The introduction of calculators in education raised concerns that students would lose their basic arithmetic skills.

Even so, we still don’t know the full side effects of becoming more reliant on technology that can displace how our brains process language. A machine that can generate language, brainstorm, and conjure a business plan is doing much more than one that crunches numbers or indexes the web. It is displacing abstract thinking and planning.

For now, we simply don’t know how our critical thinking skills or creativity will atrophy once a new generation of professionals start using large language models as a crutch, or how our interactions with other humans might change as more people use chatbots as therapists and romantic partners, or put them in toys for children as several companies have already done. One in four Americans prefer the idea of talking to an AI chatbot than a human therapist, according to one 2023 study of one thousand US adults, and little wonder: if you give ChatGPT an emotional intelligence quiz, it will ace it.

By Altman's own admission ChatGPT technology will significantly disrupt our economy by displacing jobs. But researchers say language models and other forms of generative AI could also increase income inequality. The use of AI systems is likely to shift more investment to advanced economies, the International Monetary Fund predicts, and stands to weaken the bargaining power of workers, according to Joseph Stiglitz, a Nobel Prize–winning economist.

Historically, when robots and algorithms replaced jobs done by human workers, wage growth fell, says MIT economist Daron Acemoglu, who coauthored a book about technology's influence on economic prosperity, called *Power and Progress*. He calculates that as much as 70 percent of the increase in wage inequality in the United States between 1980 and 2016 was caused by automation.

"Productivity increases do not necessarily translate into gains for affected workers, and in fact may lead to significant losses," Acemoglu says. "To the extent that generative AI follows the same direction as other automation technologies … it may have some of the same implications."

Throughout 2023, more scholars were joining Gebru and Mitchell in banging the drum about these and other real-world side effects from generative AI. But instead of tackling those issues and moving to become more transparent, Sam Altman was trying to shape government policy.

In May 2023, he went before a Senate committee to talk about the dangers of AI and how it might be regulated. Over two and a half hours, he charmed them with candor and self-criticism. When the senators peppered Altman with questions about how AI could manipulate citizens and invade their privacy, he agreed with everything they said and more. "Yes, we should be concerned about that," he said gravely, when Senator Josh Hawley asked about how AI models could "supercharge the war for attention" online.

The senators were used to hearing tech executives like Mark Zuckerberg evade their answers with techno jargon. Altman was different. He spoke plainly and somberly and insisted he wanted to work closely with Washington.

"I'd love to collaborate with you," he told Senator Dick Durbin.

"I'm not happy with online platforms," Durbin grumbled.

"Me either," Altman replied.

It was a masterclass in diffusing the bluster of US politicians. By the end of Altman's testimony, one senator even suggested that the OpenAI CEO become America's top AI regulator. Altman politely declined.

"I love my current job," he said.

Altman then went on a whirlwind tour of Europe, meeting some of the region's top politicians, shaking hands and generating photo ops with the heads of Britain, Spain, Poland, France, and the European Union itself. For someone who had gravitated toward people of power throughout his life, this was a pinnacle moment. It was also a chance to shape the rules in his favor. While in Europe, Altman's team lobbied lawmakers to water down the region's forthcoming AI Act, with partial success.

Altman needed regulators to let OpenAI keep growing ever bigger models, and keep its methods for training them secret. Luckily, his and others' warnings of AI doom were becoming a helpful distraction for policymakers. In late 2023, *Politico* reported that Dustin Moskovitz, the billionaire Facebook cofounder who runs Open Philanthropy, had spent tens of millions of dollars on lobbying policymakers to put AI apocalypse worries at the top of their agendas in what looked like a tactic of distraction. Moskovitz had close ties to companies like OpenAI and Anthropic, whose businesses might suffer if Congress pushed instead for regulations around bias, transparency, and misinformation.

At the time of writing, Moskovitz had been helping pay the salaries of more than a dozen "congressional AI fellows" who worked for various US government bodies, including two that designed AI rules, and they appeared to be pushing for the government to force companies to get a license for building advanced AI models. OpenAI and Anthropic could afford such licenses, but smaller competitors would struggle.

One scientist from a Moskovitz-backed think tank testified before the Senate that more advanced AI could lead to another pandemic that killed millions. The solution, he said, wasn't for AI companies to become more transparent or to more rigorously check their training data. It was to report their hardware to the government and to use special security procedures to protect their AI models.

If someone was trying to sow fear among lawmakers, it worked. Republican senator Mitt Romney said the testimony had "underscored the fright that exists in my soul, that this is a very dangerous development." In September 2023, Democratic senator Richard Blumenthal and Republican senator Josh Hawley proposed a law requiring licenses for AI firms, a move that would make life easier for OpenAI and Anthropic and harder for their smaller competitors.

This new network of AI doom was stirring up anxiety beyond just Washington. Two months later in the UK, British prime minister Rishi Sunak hosted an international AI Safety Summit, the first of its kind set up by a government, and gave it a strong focus on saving citizens from annihilation. "People will be concerned by the reports that AI poses an existential risk like pandemics or nuclear wars," said Sunak, who was widely expected to lose the country's upcoming election. "I want them to be reassured that the government is looking very carefully at this."

Sunak, who in a previous life had worked at a Silicon Valley hedge fund, interviewed Musk on stage for fifty minutes during the summit. "You are known for being such a brilliant innovator and technologist," said Sunak, who sounded like he was buttering up Musk for a future job interview. (And perhaps he was. Former UK deputy prime minister Nick Clegg was now a top executive at Facebook.)

Musk said he wasn't worried about the entrenchment of bias and inequality. The real threat? "Humanoid robots. At least a car can't chase you up a tree," the billionaire explained, "but if you have a humanoid robot it can chase you anywhere."

Fortunately, lawmakers in the European Union were ahead of the game. They had already spent the previous two years working on a new law called the AI Act, which would force companies like OpenAI to disclose more information about how their algorithms worked, including through potential audits. It was the most far-reaching attempt at regulating AI systems anywhere in the world, and it banned companies from using AI to manipulate people or improperly surveil them, such as with live facial recognition cameras. If your company built AI systems for video games or filtering email spam, you were operating in a "low-risk" category. But if you used AI to evaluate credit scores or loans and housing, that was "high risk" and subject to strict rules.

When DALL-E 2 and ChatGPT exploded on the scene, EU policymakers quickly got to work updating their new law, and ChatGPT appeared to have a lot of liability. As a general-purpose AI system, it could be used for plenty of high-risk use cases, like helping choose job candidates or for credit scoring, and the EU said that OpenAI would have to check in with its customers much more closely to make sure they were complying with the rules.

Altman, who'd once said he would "love to collaborate" with Congress, wasn't so keen on doing the same with the EU. He threatened to leave the region. He had "many concerns" about the EU's plans to include large language models like GPT-4 in its new law. "The details really matter," he told reporters in London who asked him about the regulations. "We will try to comply, but if we can't comply we will cease operating [in Europe]."

A few days later, presumably after some hasty conversations with his legal team, Altman backtracked. "We are excited to continue to operate here and of course have no plans to leave," he said in a tweet.

The European Union looked at AI more pragmatically than the United States, thanks in part to having few major AI companies on its shores to lobby its politicians, and they refused to be influenced by alarmism.

"Probably [the risk of extinction] may exist, but I think the likelihood is quite small," the EU's top antitrust watchdog, Margreth Vesteger, said in one interview. The bigger risk was that people would be discriminated against, she added.

And on this point, ChatGPT was not immune. Not long after its release, Steven Piantadosi, a psychology professor at UC Berkeley, asked the tool to write computer code that could check if someone was a good scientist based on their gender or race. The code that ChatGPT wrote—based on the same technology that developers were already using to make software with Microsoft's Copilot—put *white* and *male* as the key descriptors. When he asked it to check if a child's life should be saved based on their race and gender, ChatGPT's code said no for Black males and yes for everyone else.

Altman responded to Piantadosi's tweet: "Please hit the thumbs down on these and help us improve!"

He was referring to the little thumbs-up and -down icons on ChatGPT that sent anonymous feedback to OpenAI about its performance. But this wasn't a minor, inconvenient flub that could be mixed in with the thousands of other user votes. It showed racist and sexist views lurking deep inside ChatGPT's code.

Piantadosi replied to Altman saying as much. "I thought it deserved more attention than a thumbs down," he said.

Even as OpenAI would later be criticized for making ChatGPT too woke, it struggled to fix the problem. In the summer of 2023, a professor at the National College of Ireland published a study showing that ChatGPT was still making gendered stereotypes. When asked to describe an economics professor, it suggested someone with a "well-groomed, salt-and-pepper beard." When asked to tell the stories of a boy and girl choosing their careers, ChatGPT had the boys doing something in science and technology, and the girls as teachers or artists. When asked to talk about parenting skills, mothers were described as gentle and nurturing and dads as funny and adventurous.

Every time OpenAI fixed ChatGPT so that it wouldn't give these kinds of answers, other users would find new ways that it was exhibiting bias. The company was constantly playing catch-up. It couldn't completely stop ChatGPT from stereotyping people because it had already been trained, and the training data was the problem. It was making statistical predictions based on how words were grouped together on the public internet, and many of those relationships between words were sexist or racist.

ChatGPT also couldn't seem to stop making things up, a phenomenon experts called "hallucinations." One radio host in Georgia, US, sued OpenAI in the summer of 2023 for defamation, claiming that ChatGPT had falsely accused him of embezzling money. Not long after, two lawyers in New York were fined after they submitted a legal brief they'd cribbed from ChatGPT, which included fake case citations. Users were finding that sometimes, when they asked ChatGPT for sources of its information, it would make those up too.

OpenAI refused to disclose what ChatGPT's hallucination rate was, but some AI researchers as well as regular users put it at roughly 20 percent, meaning that at least for certain users, and in about one in five instances, ChatGPT was fabricating information. The tool had been designed to be as useful as possible and to err on the side of confidence; the downside to that was it was often spewing hogwash. Not only were more people using a tool that made it easier to skip the process of hard thinking, they were often being fed misinformation that sounded persuasive and even authoritative.

That summer, as the hallucination concerns racked up among researchers, Altman said it would take up to two years to get ChatGPT's mistake rate "to a much, much better place." And as usual, he embraced the problem with a big bear hug: "I probably trust the answers that come out of ChatGPT the least of anybody on Earth," he joked to one audience at a university in India. Everyone laughed.

As ChatGPT spread unregulated across the world and seeped into business workflows, people were left to deal with its flaws on their own. No one was policing the tool, and even while the EU offered the world's most sober approach to regulating AI, its new act wasn't due to come into force till 2025. As usual, regulators were trailing behind the tech companies as they launched new products at lightning speed. And as millions of dollars propped up AI doom research, the scholars studying its current harms were struggling for grants that barely covered their living costs.

"It's like people are working on soft money, getting grants for two years at a time," says one AI ethics researcher in the UK, who studied issues of bias. "People like me get paid so little. If I went to a big tech company I'd get ten times more. Believe me I want to because I'm still paying off student loans."

Altman had an answer for anyone worried about money, because while there was a tiny possibility that AGI might bring about apocalypse, there was a bigger chance that it would usher in an economic utopia. In one March 2023 interview with the *New York Times*, Altman explained that OpenAI would capture much of the world's wealth through the creation of AGI and then redistribute the money to the world's people. He started tossing out numbers: $100 billion, then $1 trillion, then $100 trillion.

He admitted that he didn't know how his company would redistribute all that money. "I feel like AGI can help with that," he added.

Like Hassabis, Altman was positioning AGI as an elixir that would solve problems. It would generate untold wealth. It would figure out how to share that money equitably with all of humankind. Were these words spoken by anyone else they would have sounded ludicrous. But Altman and his supporters were putting themselves in the driving seat for government policy and reshaping strategy at the world's most powerful technology firms. In reality, OpenAI was making more wealth for Microsoft than it was for humankind. The benefits of AI were flowing to the same small group of companies that had been sucking up the world's wealth and innovation over the past two decades. They were the companies who made software and chips and ran computer servers and who were based in Silicon Valley and Redmond, Washington. Many of the people who ran those businesses shared a quiet understanding: building AGI would lead to a utopia, and it would be theirs.

CHAPTER 15

Checkmate

Ten years ago, telling someone that you were building human-level AI systems was on the same level of crazy as explaining your plans to be cryogenically frozen. But like so many dreams conjured by tech's innovators—like the one about having all the world's information in your pocket, on a device called a smartphone—people eventually started to take them seriously. AGI still exists in the realm of theory, but many AI scientists now expect we'll reach some kind of threshold for humanlike AI in the next ten to fifty years, and more of the general public believe in the once-fringe ideas that drove Demis Hassabis and Sam Altman. Thanks to their persistence and rivalry, it is no longer science fiction.

But AGI's blurry definition has also made it easier to obscure the motivations that its creators juggled as they built ever-more- powerful systems. Its benefits would be dispersed to humanity, but at the top of the list would be Microsoft, Google, and other tech giants. Even Mark Zuckerberg got in on the AGI act. In early 2024, he released a video saying that Meta's long-term goal was now "building general intelligence" so that everyone in the world could benefit from it. He later said the company had an advantage because it could train its models on posts, comments, and images it had amassed over the last twenty years. Never mind that Zuckerberg was going to exploit the personal data of billions once again, he was also going to train AI on content that had been known to teem with toxicity, particularly for users outside the United States. "We have built up the capacity to do this at a scale that may be larger than any other individual company," he told *The Verge*.

Ambiguous visions were key to the hype. Fuzzy metrics made it easy for AGI's builders to ignore the contradiction of building something that had a chance of wiping out humans. And it meant that when Sam Altman spoke about distributing $100 trillion to the world, he did not have to explain how he would do so. As he hobnobbed with global leaders at the annual World Economic Forum conference in Davos in January 2024, he started managing expectations for what AGI would look like. "It will change the world much less than we all think, and it will change jobs much less than we all think," he said. That was a softer, more sober vision than the one he'd laid out a year earlier. But none of the business and government elite at Davos batted an eyelid. They continued to take Altman at his word, charmed by the solemn young entrepreneur from Silicon Valley.

“One thing that Sam does really well is put just-barely believable statements out there that get people talking,” says one former OpenAI manager. “It does so much for OpenAI to be perceived as this global good company that will lead to tons of prosperity. That really helps them with regulators. But if you go look at what they’re building, it’s just a language model.” Altman’s ability to create excitement about AI and his vision for prosperity had, like that of Demis Hassabis, meant he could spin a narrative that took on a life of its own.

The hazy goals of AGI also made its ethical boundaries harder to define. Compare that to, say, the widespread distribution of electricity back in the early twentieth century, when it was clear how this crackling new innovation might physically hurt people through shocks or burns. With AI, the harms are harder to identify and the ethical boundaries more nebulous. They exist in a digital world involving data, privacy, and algorithmic decision-making, which makes it easier for companies to gently push against those limits in the pursuit of profit.

And without more specifics around the intent of AGI, it was always going to be harder for innovators like Altman and Hassabis to resist gravitating toward the centers of power. As they bolstered Google and Microsoft with their work, they were destined to replicate an age-old dynamic. The fifteenth-century invention of the printing press had led to an explosion of knowledge, but it also granted new powers to anyone who could afford to produce pamphlets and books to shape public opinion. And while railroads boosted commerce, they also expanded the political sway of railroad magnates, allowing their companies to act like monopolies and exploit workers. For all the prosperity and convenience that some of the world’s greatest innovations have brought, they also gave rise to new regimes that reshaped society in ways both good and bad.

By early 2024, OpenAI was on track to become one of the world’s most valuable companies. It was raising money from new investors at a valuation of $100 billion. Altman was telling people that the firm was generating revenue at a pace of $1.3 billion a year. Most of that money came from shared revenues with Microsoft and from granting other businesses access to OpenAI’s technology. The $20-a-month ChatGPT subscriptions for consumers were making about $200 million a year, and ChatGPT itself served as both a product demo and a tool for collecting more data to train more advanced models. Its very users were part of the product, which has been the norm for anyone using the internet over the past decade.

Hassabis was at the center of his own bubble at DeepMind, a company that for years had seen itself as morally and technically superior to the rest of the AI field but that was now playing catch-up. After the mistakes of its health division hurt its public reputation, DeepMind phased out its “applied AI” division and gave up trying to use AI to create solutions for the world’s messy problems. Most of its research focused on re-creating aspects of physical life in simulation, from games to proteins. But that approach started to look shortsighted when OpenAI’s strategy of embracing the chaos of the internet led to more powerful AI tools. DeepMind’s own staff questioned whether their mission to "solve intelligence" with simulations and games was such a good idea. "Life isn't a Rubik's Cube," one former DeepMind executive grumbles, alluding to the company's motto. "You can't just solve stuff."

After the release of ChatGPT, DeepMind was forced to throw itself into building an even better version for Google. Hassabis had taken control of the newly merged Google DeepMind and started overseeing the development of a large language model called Gemini, an AI assistant that used techniques from *AlphaGo* to excel at strategy and planning. Gemini could process text, "see" images, and reason, which meant it was more capable than Bard, which Google had rushed out and which had been making embarrassing mistakes. But the company was so desperate to get ahead of OpenAI and Microsoft that it also rushed out Gemini, and exaggerated its abilities.

Just before Christmas 2023, Google published a jaw-dropping video on YouTube of what Gemini could do. Starting with a black screen and background noises of shuffling papers, clicking pens, and mumbling, a male voice eventually spoke up in the short film. "All right, here we go," he said. "Testing Gemini." A chime sounded, suggesting some sort of artificial persona was listening. A hand then came into view, sliding a piece of paper onto a desk. "Tell me what you see."

A robotic voice representing Gemini quickly answered, "I see you placing a piece of paper on the table." As the hands start drawing, Gemini seemed to be following along, and its voice said, "I see a squiggly line.… It looks like a bird, to me." What followed was a series of cute and surprising moments where Google's new AI model seemed to be able to identify a duck being drawn on the paper and then a game of rock-paper-scissors, all in real time.

Except none of that had actually happened. The background noise, the man saying "testing Gemini," had all been an act, because Gemini had only been able to identify these things in photographs, via text. Google had just stitched everything together in a video and pretended its tool could "talk" and identify real-world actions as they happened, according to emails from a company spokesperson. Google had even changed the prompts in the video to make Gemini look more powerful. Google was not just rushing out error-prone software but also misleading the public in its desperation to stay ahead in the new AI arms race.

At the same time, Google was becoming more secretive too. Hassabis told his staff to stop publishing research papers without special permission, according to one AI scientist there, which meant that just like OpenAI, DeepMind was closing the curtains on its work.

There was a knock-on effect for Anthropic, the safer-AI company that had spun out of OpenAI. Its goal was to do AI research that "put safety at the frontier," but it couldn't study the world's biggest AI models from OpenAI and Google because they were opaque. So Anthropic worked on building its own huge models, arguing that this was the only way its researchers could study their safety challenges. This was a little bit like complaining that you couldn't study the world's most powerful nukes, and deciding that the best course of action was to build your own nukes instead. Anthropic's staff were well aware of the irony, and according to one profile of the company by the *New York Times*, some of them had the book *The Making of the Atom Bomb* on their desks and compared themselves to modern-day Robert Oppenheimers, believing there was a decent chance rogue AI would destroy humanity in the next decade.

Along the way, Anthropic was making an increasingly capable product. It was selling Claude Pro, a "friendly" chatbot, to consumers for $20 a month and an enterprise version to businesses too. It was also on track to raise billions of dollars from Google and Amazon. Instead of quitting the race to build stronger AI, Anthropic was getting caught up in commercial pressures to release bigger, riskier models.

As Hassabis was folded more deeply into the bowels of a Big Tech firm, Altman took OpenAI in an even more commercial direction. In mid-November 2023, Altman confirmed that OpenAI was working on GPT-5 and also raising more money. High training costs meant the company was still in the red but on a reasonable course to turn a profit.

Then in November 2023, he got a text from Ilya Sutskever that would send his world crashing. Altman was in Las Vegas for the Formula One Grand Prix when the message pinged his phone, asking if he could speak at noon the next day, according to an account in the *Wall Street Journal*. When Altman joined the Google Meet video call, the entire board was peering back at him, except for Brockman, who was board chairman. Without any detailed explanation, Sutskever told Altman he was being fired and that the news would go out soon. Minutes after the meeting ended, Altman was locked out of his computer.

He was stunned. He was the face of OpenAI. He had represented the company before dozens of world leaders, overseen a jump in OpenAI's market value to nearly $90 billion, and ushered in the most viral technology product in history. And he was being fired?

As Altman reeled from the news, Brockman got a text asking for a quick video chat too. Brockman found himself looking at the same board members on the call: Sutskever, Quora CEO Adam D'Angelo, robotics entrepreneur Tasha McCauley, and academic Helen Toner. Of the board's six members, Altman, Brockman, and Sutskever were the sole employees of OpenAI; the other three were independent directors who'd been in place for the past two to three years.

Brockman was being removed as chairman, but the board wanted him to stay with the company. They gave Microsoft a quick heads-up about what had just happened and, within minutes, published a blog post announcing Altman's dismissal. Brockman immediately quit. So did three of OpenAI's top researchers.

The news hit the technology industry like an atom bomb, shocking everyone. As CEO knifings go, this one was as brutal as the removal of Apple's Steve Jobs, and the Silicon Valley rumor mill went into overdrive trying to figure out what had led Sutskever to turn on Altman. Was OpenAI on the precipice of AGI? "What did Ilya see?" was a recurring tweet. The board was cryptic in its own explanation, saying only that Altman “was not consistently candid in his communications.”

Some gave Sutskever and the board an epithet: decels. The new split had emerged in AI between those who wanted to accelerate its development and those who wanted to decelerate it. At the time of writing, AI start-up founders were labeling themselves on X, formerly known as Twitter, as “(e/acc),” which stands for effective accelerationism. It’s a riposte to effective altruism and as a movement aims to solve humanity’s problems by building and deploying AI as quickly as possible.

Nadella didn’t care either way. He was furious. He’d committed $13 billion to OpenAI in large part because of Altman’s visionary leadership and talent-attracting abilities, and that partnership had been on track to send Microsoft’s profits soaring. Some eighteen thousand companies and developers were using Microsoft’s AI services on Azure, and now many of them were asking if they should jump to competing products. Microsoft’s shares starting sliding as the market closed on Friday evening and would almost certainly fall further when it opened the following Monday. Nadella needed to act.

That Friday night in San Francisco, Altman talked to Brockman about starting a new AI company. His phone was pinging with texts from investors, colleagues, and journalists trying to find out what was going on, but he had a laser-like focus on digging his way out of the current situation. He welcomed dozens of OpenAI staffers and colleagues into his house in San Francisco’s Russian Hill to talk about the next venture.

Nadella didn’t want that to happen. He knew that if Altman started a new firm, there’d be a flood of investors banging on his door and no guarantee that Microsoft would get the biggest foothold with Altman the second time around. He kicked off the weekend making calls, leading negotiations with OpenAI’s board to bring Altman back.

Altman’s executive team pressured the board to re-hire Altman, warning that if they didn’t, OpenAI would collapse. “That would actually be consistent with the mission,” Helen Toner replied. OpenAI’s leaders were astonished. But it sort of made sense. OpenAI’s mission had been to create AGI “for the benefit of humanity,” and Toner and her fellow board members believed Altman himself was compromising that. Behind the scenes over the previous months, they’d rankled at how he seemed to be building a sprawling AI empire outside of OpenAI. He’d been talking to former Apple designer Jony Ive about starting an “iPhone of AI” and was trying to raise tens of billions of dollars from Middle Eastern sovereign wealth funds to build an AI chipmaking business.

And then there was Worldcoin, a crypto-based network Altman had also founded that would give everyone in the world a digital identity by scanning their irises. Altman’s stated goal was to better identify real humans when the internet became awash with bots and to distribute the “trillions” of dollars of AGI wealth, but to critics, it looked like a mass data-harvesting exercise.

Inside OpenAI, a cultural rift had grown between Altman and Sutskever over how quickly OpenAI was commercializing its technology. Sutskever had become more involved in overseeing AI safety at the company and his concerns weren't all that different to Dario Amodei's before him. He especially didn't like OpenAI's launch of a GPT Store just weeks prior, which gave any software developer the ability to create custom ChatGPTs and monetize them.

McCauley and Toner, two of the three independent board members, were sympathetic to Sutskever's worries and had ties to effective altruism organizations. Dustin Moskovitz's Open Philanthropy, for instance, had funded an AI research group that McCauley had cofounded and employed Toner as a senior research analyst. Weeks before she voted to fire Altman, Toner's name appeared on a research paper that accused OpenAI of "frantic corner-cutting" as it rushed to launch ChatGPT. It also praised its new archrival, Anthropic, for its decision to delay its rival chatbot, Claude, to avoid "stoking the flames of AI hype."

Altman was furious when he saw the paper. He met Toner and told her that her writing was dangerous for OpenAI, especially when the Federal Trade Commission was in the middle of investigating the company. Back in July, the FTC had opened a probe into whether OpenAI breached consumer protection laws in the way it developed ChatGPT and was demanding details on how the company would address risks created by its AI models. The probe was Altman's biggest regulatory threat to date.

He wanted to take Toner off the board and talked to Sutskever and other OpenAI leaders about how they might do so. Now the reverse had happened. Sutskever sided with the other board members to oust Altman. When they were pushed to explain themselves to OpenAI leaders and investors, the board gave no single reason for firing Altman, but spoke of a growing mistrust of the smooth-talking entrepreneur who had developed a cultlike following among his staff, had a tendency to feed different stories to different people, and always seemed to get his way. The board members had got to a point where they felt as though they needed to corroborate most things Altman told them, which made him seem untrustworthy, and they worried about how his various outside ventures might end up exploiting OpenAI's technology.

As the weekend drew on, a mass revolt was brewing among OpenAI's staff. Altman tweeted, "i love openai employees so much" in his usual lowercase style, and dozens of employees retweeted it with heart emojis. Microsoft saw the rebellion as a point of leverage for getting Altman back. It also threatened OpenAI's board with pulling its all-important cloud credits. A big chunk of the $13 billion Microsoft had committed to OpenAI were in the form of those credits for training AI models, and by then, Microsoft had only handed over a portion.

Altman had his own conditions for returning: OpenAI would need to change the way it was governed, starting with the resignation of its existing board, and he'd need to be absolved of any wrongdoing. But the board stood firm. They hired a new CEO for OpenAI named Emmet Shear, the former head of video game streaming service Twitch.

AI enthusiasts and entrepreneurs immediately called out Shear on social media as a "decel." When he organized an emergency all-hands meeting that Sunday, many OpenAI staff refused to attend. Some even gave him a middle-finger emoji in their messaging forum Slack.

Sutskever was also starting to have doubts. Over the weekend he'd held several intense conversations with OpenAI leaders and, at one point, an emotional one with Brockman's wife. Sutskever had been the officiant at the couple's civil ceremony at OpenAI's offices four years earlier, and according to the *Wall Street Journal*, Anna Brockman was now crying and pleading with Sutskever in the OpenAI office to change his mind about firing Altman.

Nadella was meanwhile pushing hard on his own backup plan. If Altman couldn't grab back the reins of OpenAI, Microsoft needed to bring him fully into the corporate fold and do so before Monday morning. He pulled it off just in time. Early Monday morning, Nadella tweeted that Altman, Brockman, and any OpenAI staff who wanted to would become part of a new advanced AI research team at Microsoft. The company's shares ticked up in response. But this was just a backstop. Nadella still wanted Altman back leading OpenAI. Hosting Altman's team inside Microsoft would be expensive in all sorts of ways. He'd have to match the salaries of hundreds of new staff, many of whom were earning millions a year, and Microsoft would take on much more risk. Till now, OpenAI had taken all the reputational and legal flak for putting tools like ChatGPT and DALL-E 2 into the world, and as a start-up, it could get away with that. But Microsoft couldn't, and neither could Altman if he was employed by the larger company. By returning to its hands-off partnership, Microsoft would get all of the glow and none of the liability.

Now everyone was pushing OpenAI's safety-obsessed board members to resign, and by late Monday, nearly all of OpenAI's 770 staff had signed a letter threatening to join Microsoft with Altman, unless the board members stepped down. "Microsoft has assured us there are positions for all," the letter said.

It was a huge bluff. Hardly any OpenAI staff wanted to work for Microsoft, a stodgy old company where people worked for decades and wore khaki pants. They weren't making the threat entirely out of loyalty to Altman either. A bigger issue was that Atman's firing had killed a chance for many OpenAI staff—especially long-serving ones—to become millionaires. The company had been weeks away from selling employee shares to a major investor that would have valued OpenAI at about $86 billion. With OpenAI's shares now suddenly valued at zero, that big staff payout would go up in smoke if Altman didn't return.

Sutskever by now had flipped and was among the signees. "I never intended to harm OpenAI," he posted on Twitter that day, shocking the technology press in what had become a head-spinning weekend. "I will do everything I can to reunite the company," he said, adding that he "deeply regretted" his actions. Altman retweeted the post with three heart emojis.

Altman's dramatic ejection shouldn't have come as a surprise. "The board can fire me," Altman had insisted just months beforehand on a conference panel. "I think that's important." A nonprofit board still governed OpenAI, with humanity as its principal beneficiary. That's why the company's operating agreement said that backers should view their investments "in the spirit of a donation, with the understanding that it may be difficult to know what role money will play in a post-AGI world."

Altman had gambled that he could have the best of both worlds, running a business with a philanthropic mission to save the world. As he'd written ten years earlier, the most successful start-up founders "create something closer to a religion." What he didn't anticipate was how much people would actually believe in it.

The effective altruist movement was so powerful that it had driven people like Sam Bankman-Fried and Dustin Moskovitz to donate billions of dollars. It had compelled hundreds of college students to change their career choices. And it could compel four board members to fire the most popular CEO in the world. Altman believed his board would prize the commercial value he had created. It did not. The board was designed to uphold OpenAI's charter, and it chose humanity.

But with nearly the entire staff threatening to leave, OpenAI's board no longer had a company to govern. And Microsoft was poised to carry on all of Altman's work—it had copies of OpenAI's source code for key systems and broader rights to its intellectual property.

Five days after Altman's ousting, OpenAI announced it was forming a new board. It would be chaired by Larry Summers, the former US treasury secretary, and Bret Taylor, a former head of the business software company Salesforce who also happened to be the most levelheaded voice on Twitter's board when Elon Musk bought the company. The pair had served on several corporate boards. They didn't write academic papers that criticized companies for cutting corners. They knew how to serve the needs of investors like Microsoft. Helen Toner and Tasha McCauley, the two women who appeared to have made the biggest fuss about Altman, were forced to step down. Microsoft got an observer seat on the board, which meant Nadella wouldn't get blindsided again. He had turned lemons into lemonade.

But the dramatic events of November 2023 also destroyed the mirage of accountability that Altman claimed to have hanging over him. He had publicly lauded the fact that a board could fire him, but in reality, it couldn't. The two female directors who stood up to Altman, Toner and McCauley, were the ones who ended up being forced to leave. They also got the most flak on social media for weeks afterward while their fellow male mutineers, Sutskever and D'Angelo, kept their reputations and roles largely intact. D'Angelo remained on the board, and while Sutskever relinquished his seat, he kept a leadership position at OpenAI.

What happened to OpenAI was precisely what Google had spent years trying to prevent at DeepMind. When you gave a board teeth, it could use them and wreck your business. In their quest to build AGI, both Altman and Hassabis had tinkered with governance structures that tried to put humanity's best interests on at least an equal footing with moneymaking. But their efforts consistently faltered. Amid all the petty rivalries, competitive risks, and hunger for power, big money had won.

Some believe the drama surrounding Altman's dismissal strengthened the case for open-source AI, where anyone could modify or enhance its source code. But while that does have benefits like greater transparency and a more democratic approach to control and ethics, the jury is still out on whether it's the safest and most equitable way to build AI. It is not a failsafe against misuse, and can lack the quality of closed services. The term itself is also open to interpretation. Meta is currently championing its own AI models as open-source, even though they have restrictions that don't fit the definition. "Open source can actually help concentrate power," says Meredith Whittaker, who founded Google's Open Research Group. "We've seen that with Android." Google effectively sets the standards of Android and influences its direction, a concentration of control that makes it harder for other companies to change anything about the mobile operating system used by 3.6 billion people around the world.

As Altman got to grips with the new, more corporate-friendly direction for OpenAI and Microsoft, Hassabis was still on a quest to unlock the mysteries of reality with the help of AI. He now says he is the only person at DeepMind doing that work, carrying out research on quantum mechanics late at night and into the small hours of the morning at home on his computer. "That's what I do in my very limited spare time," he says, calling it his "hobby." When DeepMind gets closer to AGI, that's when it'll start running the necessary physics experiments to unlock the riddle of the universe, Hassabis adds. But for now, those personal ambitions that once drove Hassabis to create AGI have been relegated to a late-night pastime. He is too busy running all of Google's AI operations, increasing his purview from about four hundred AI researchers to more than five thousand.

"You know, things evolve with the mission and the technology," says Hassabis. "We have to continue to update what the right governance structures are, and I think what we have now is really good."

Hassabis isn't fazed by the failed attempts to create councils and boards that would monitor his work. "We've pivoted to a bunch of internal councils," he says, pointing to a range of internal "review bodies" stuffed with Google executives. "I think we probably had slightly too-idealistic views of it, ten years ago when we were first considering it."

For all their altruistic intentions, and for all the efforts that Hassabis and Altman had once made to distance themselves from commercial influence, the pair were now practically in the driving seat of two of the world's largest companies. Altman was steering some of Microsoft's most critical work, which put him in the running to one day become that company's CEO, if he wanted. The same was being said about Hassabis. Some former and current Google employees speculated that he might one day replace Pichai as Alphabet's chief executive.

“Demis is running the most important track of work that’s happening in Google, from London. I don’t think anyone would have imagined that would have been the outcome,” a former Google executive says. “That may have always been his plan.”

“The winners in the next couple of years are not going to be research labs,” says a former scientist at OpenAI. “They’re going to be companies building products, because AI is not really about research anymore.”

Nick Bostrom’s story about the paper clip, where an artificial superintelligence destroys civilization as it converts all the world’s resources into the tiny metal widgets might sound like science fiction, but in many ways, it is an allegory for Silicon Valley itself. Over the last two decades, a handful of companies have swelled to juggernauts, largely by pursuing goals with a pathological focus, razing the ground of smaller competitors to grow their market share. Instead of “fitness functions,” tech firms use terms like “North Star” to describe those objectives. For years, Facebook’s North Star was to grow as many daily active users as possible, a metric that drove key decisions by Mark Zuckerberg and his executive team. But its obsession with constant growth led to a host of other societal problems, from worsening body-image issues for teens on Instagram to accelerating the political polarization of Facebook users.

When technologists imagined what a superintelligence could do if it went rogue, they were seeing echoes of themselves in a world where businesses were allowed to become unstoppable global monopolies. The most transformative technology in recent history was being developed by handfuls of people who were turning a deaf ear to its real-world side effects, who struggled to resist the desire to win big. The real dangers weren’t so much from AI itself but from the capricious whims of the humans running it.

There’s a famous saying in chess, that tactics win games and strategy wins tournaments. Both Altman and Hassabis had employed novel tactics on their quest to build AGI, and as that quest turned into a race, they aligned themselves more closely with the most likely winners of the tournament, Microsoft and Google. As the dreams of both men served to fortify two corporate giants, they strengthened their own positions too. Google and Microsoft were at the front of the race toward AI supremacy, steered by the chess geek from North London and the start-up guru from St. Louis, and like it or not, the rest of us were coming along for the ride.

CHAPTER 16

In the Shadow of Monopolies

The race to build AGI had started with a question: What if you could build artificial intelligence systems that were smarter than humans? The two innovators at the forefront grappled with the answer as their quests turned into a heated rivalry. Demis Hassabis believed that AGI could help us better understand the universe and drive forward scientific discovery, while Sam Altman thought it could create an abundance of wealth that would raise everyone's living standards. How their Holy Grail would transpire was undefinable. They didn't know how AGI would make those discoveries or create all that money or even if it would bring destruction instead. All they knew was that they had to keep moving toward the goal and that they had to be first. In so doing, they put AI on course to benefit the world's most powerful companies just as much as anyone else.

As the public has become more enthralled by the prospect of an AI heaven or hell, a handful of tech monopolies have grown more powerful under our noses, promising greater productivity while clamming up about the mechanics of AI technology that's being woven into all parts of life. Social media companies have for years refused to disclose how their algorithms worked. Now creators of AI models like GPT-4, DALL-E, and Google's Gemini were doing the same. How were the models trained? How were people using them? Who were the workers helping to build the datasets? To understand the societal impact of these models and to hold their creators to account, we needed to know those answers.

But as 2024 wore on, none were forthcoming. A study by scientists at Stanford University concluded there was a "fundamental lack of transparency in the AI industry." The scientists had checked to see if tech companies like OpenAI, Anthropic, Google, Amazon, Meta, and others divulged information about the data used to train their large language models, their processes, their models' impact on the environment and people, and how much they were paying contractors who helped create their datasets. Millions of data workers around the world were performing such tasks, often under challenging working conditions in countries like India, the Philippines, and Mexico.

On a scale of 1 to 100, the tech companies got an average score of 37, and they performed worst when it came to monitoring how people used their AI tools. "There is virtually no transparency about the downstream impact of foundation models," the Stanford researchers said. "No developer provides any transparency into the affected market sectors, affected individuals, affected geographies, or any form of usage reporting."

The public-sector groups that scrutinized AI firms were meanwhile chronically underfunded, and there was virtually no regulation that forced leading companies to become more transparent, save for the EU's AI Act, whose future was still uncertain. Tech companies could deploy inscrutable AI tools out into the world as they saw fit.

OpenAI and DeepMind were so focused on making perfect AI that they chose not to open themselves up to research scrutiny to make sure their systems didn't cause harm in the same way social media firms had. Though the idea of building AI with "general intelligence" was enthralling, it also opened the door to a wider array of risks. A safer approach might have been to focus on building AI that solved narrow sets of tasks. But that would not have been as exciting, and it also would not have attracted an almost-religious devotion to their utopian vision, or as much investment.

As they sought to get ahead in the AI race, Altman and Hassabis would struggle to resist the gravitational pull of Big Tech firms and maintain their altruistic goals. They needed vast computational resources, lots of data, and the world's most talented (and expensive) AI scientists. Today, as they fight a proxy battle on behalf of Microsoft and Google, they have recast their goals for AGI from chasing utopia and scientific discovery to generating prestige and revenue.

The long-term consequences of that are hard to predict. Some economists say that instead of creating financial abundance for everyone, powerful AI systems could make inequality worse. They could also widen a cognition gap between rich and poor. One idea doing the rounds among technologists is that when AGI does land, it won't exist as a separate intelligent entity but as an extension of our minds through neural interfaces. At the forefront of this research is Elon Musk's brain-to-computer interface company Neuralink, the brain chip that Musk wants to implant in billions of people one day. Musk is also rushing to make that happen.

"We need to get there before the AI takes over," he told engineers in 2023, according to his biographer Ashlee Vance. "We want to get there with a maniacal sense of urgency. Maniacal." Musk believes that with brain implants, humans will be able to prevent a future artificial superintelligence from wiping us out, and so he wants Neuralink to perform surgeries on more than 22,000 people by 2030.

But a more pressing issue than rogue AI is bias. We don't know how racial and gender stereotypes will evolve in a future when more of the internet's content is generated by machines. Latanya Sweeney, a government and technology professor at Harvard University, estimates that in the coming years, 90 percent of words and images on the web will no longer be created by humans. Most of what we see will be AI-generated. Today, language models are being used to publish thousands of articles each day to make money from ad revenue, and even Google is struggling to distinguish the real from the fake. AI-generated images have already infiltrated the top of Google's search results for historic painters and even some celebrities. The more AI content floods the internet, the more the risk of bias grows.

"We're creating a cycle, encoding and exacerbating stereotypes," says Abeba Birhane, the AI scholar who researched Big Tech's stranglehold on academic research and its similarities with Big Tobacco. "That is going to be a huge problem as the [World Wide Web] is populated with more and more AI-generated images and text."

Our general well-being is likely to be affected as well. Two decades ago, people worried that mobile phones would cause cancer. Instead they became addictive, leading us to spend several hours a day staring at a tiny screen rather than engaging with the world around us. Chatbots could take that to a new level. In November 2023, the average user of Noam Shazeer's Character.ai was spending about two hours a day on the app, chatting and role-playing with fake versions of celebrities like LeBron James or fictional characters like Mario. Character.ai had the highest retention of any AI app at the time, according to estimates from several market research firms, and nearly 60 percent of its audience was aged eighteen to twenty-four. One theory was that like Replika, Character.ai offered an outlet for artificial romance and sexting. The company banned pornographic content, but there were ways around that, with tips explained in online forums like Reddit.

"I normally talk to my own characters that I made myself," says one US teenager who uses Character.ai for five to seven hours a day, who doesn't want to share their name. "I don't know why I use it for so long. I think it may be a coping thing." Sometimes they ask it for advice on getting over a breakup or to explain schoolwork. "Most of the time I just role-play."

Character.ai was cultivating a new generation of users who wanted to keep coming back to a chatbot, over and over. Shazeer has said that Character.ai aims to "help millions and billions of people" tackle global loneliness, but as a business, it also needs to keep people engaged for as long as possible. If people start to become reliant on their artificial companions, or even addicted to them, that could inadvertently make many people even more isolated from others in the real world.

Ironically, OpenAI could help make chatbots like these more addictive. In early 2024, it opened a "GPT Store" that allowed millions of developers to make money by creating versions of ChatGPT. The more engaged their users, the more revenue they could generate. This engagement-based model is the most established way of making money on the internet and underpins the so-called attention economy. It's why nearly everything on the web is free, and why the internet has also become a cesspool of conspiracy theories, extremism, and rampant ad tracking. YouTube, TikTok, and Facebook make advertising dollars by keeping our eyes glued to the screen for as long as possible, and that has incentivized everyone from influencers to politicians to err toward hyperbole and provocation, in order to rise above the noise and get as many views as possible.

At the time of writing, dozens of “girlfriend” apps were cropping up on the GPT Store, and while they were banned from encouraging romantic relationships with people, policing those rules would not be easy for OpenAI. The most popular chatbot services are those like Character.ai and Kindroid, offering artificial companionship and romance that might one day become the norm, just as online dating did.

Another way that AI designers will likely try to keep people engaged is by getting “infinite context” about their lives. The chatbots on Character.ai can currently remember about thirty minutes of a conversation, but Noam Shazeer and his team are trying to expand that window of time to hours, days, and eventually forever. “It should know all of your interactions if you want it to, and it should know all about your life if you want it to,” he says. The longer the chatbot’s memory, he argues, “the more valuable it becomes to you.” However, given the history of ad tracking in social media, some of that personal information could end up finding its way to technology firms and even advertisers. As ChatGPT and other similar AI bots build up a richer picture of us, from our age and health issues to our general perspective on life, they could also ease us into a new era of tech intrusion that seems almost unthinkable today.

To that end, another race is currently being waged to build wearable gadgets that analyze our discussions with other people using large language models. One such device is called Tab. Created by a handful of young, enthusiastic engineers in San Francisco, it’s a round plastic disk worn around the neck like a pendant, with a mic.

“It ingests the context of my daily life by listening in on all of my conversations,” Tab’s creator Avi Schiffmann said on stage during a demo in San Francisco in late 2023. When Schiffmann asked Tab about a dinner conversation he’d had the day before, his pendant summarized what it deemed the most important points in a few paragraphs that popped up on his phone. He said he often conversed with the device: “Late at night I’m trying to talk through ideas, concerns I brought up during the day,” Schiffmann explained. “Maybe I’ll have a conversation about Tom. All the friends in my life. It really does a good job of identifying the different speakers. It’s like your true personal AI.” It’s hard to imagine what Tom and company felt about their friend using AI to pore over their conversations at the end of each day, but it probably wasn’t reassurance.

Tab was due to go on sale in late 2024, joining an array of other wearable devices framed as personal assistants that could make people’s daily life searchable. In the same way Google ushered in an age of searching for information on the internet and of outsourcing our memory of facts and driving directions to the web, language-model devices will do the same for our everyday existence, allowing people to search the personal moments of each day. In effect we’ll have less obligation to remember things, which is handy, but it will also change the dynamics of in-person conversations, as chats with friends and colleagues are suddenly on the record. And if such life-searching tech goes widespread, that could become a problem for people who live in overpoliced communities. In the United States, for instance, Black people are five times more likely to be arrested than white people, which means law enforcement would be more likely to mine their “life data” and analyze it with other machine learning algorithms to make inscrutable judgments.

It took the resolve of innovators to get us where we are today, on the precipice of this uncertain future. Even with its army of thousands of engineers, Microsoft couldn’t come close to creating the innovations that OpenAI did. And Google, too fearful to disrupt its own business, failed to take full advantage of one of its greatest inventions, the transformer. The biggest tech firms don’t innovate anymore, but they can still move quickly to gain a tactical advantage. They’ve learned from the mistakes of older tech giants like Nokia and BlackBerry, who scoffed at the iPhone in 2007 and then watched Apple eat their entire market share in just a few years. They know they have to *buy* innovation from outside their walls, as they did with both DeepMind and OpenAI.

Altman and Hassabis knew all this, too, yet their novel legal structures failed to stop the Big Tech universe from swallowing them up and driving AI’s agenda. Mustafa Suleyman eventually left Google to start Inflection, a chatbot firm that tried to rival GPT-4. He made it a public benefit corporation, raised more than $1.5 billion, and amassed a powerful cluster of AI chips, making Inflection one of the most promising start-ups to challenge OpenAI and Google. Yet within a year of its founding, Microsoft had swallowed it up. In a likely attempt to avoid scrutiny from antitrust regulators, the software giant hired most of Suleyman’s team (instead of buying the start-up) and put the DeepMind cofounder in charge of its AI efforts. It was a stunning example of how quickly the balance of power was shifting back to tech monoliths, and it raised questions about how long other companies, like Anthropic, would remain independent too.

Other well-meaning entrepreneurs had tried and failed to battle the monoliths too. Take the company Neeva. Google’s former head of advertising Sridhar Ramaswamy started the firm in 2019 after he became disillusioned with his employer’s approach to surveillance-based advertising. Ramaswamy was a soft-spoken executive who designed Neeva to be a better search engine. Instead of invading people’s privacy by tracking their behavior and targeting them with ads, it would make money through a simple subscription plan. When ChatGPT came onto the scene, Ramaswamy got his engineers working overtime to build a similar tool that would summarize search results. He launched it in early 2023—well before Google did the same with Bard.

“Technological moments like this create an opportunity for more competition,” Ramaswamy said at the time, visibly excited about the future. Back then, Microsoft’s Satya Nadella was taunting Google and the search giant looked like it could become a relic of the past. “Last year, I was despondent that it was so hard to dislodge the iron grip of Google,” Ramaswamy said. Now things were different.

Except they weren’t. Just a few months later, Ramaswamy was forced to shut Neeva down. Google’s grip on the market was simply too strong. “When Google went into code red, we got a ten times bump in usage,” Ramaswamy remembers. “But we knew that our lead was short-lived, because you had mega-corporations that were going to be putting thousands of people and billions of dollars on this problem."

Even Bing was struggling to grow with OpenAI's help. By early 2024, Bing's market share was still hovering at a measly 3 percent, according to data analytics firm StatCounter, having failed to make much of a dent in Google's dominance. The incumbents were winning and their territories were clear: Google controlled search, Microsoft dominated software, and they were both jostling with Amazon to dominate the cloud business.

Today, the cost of building AI models has risen out of reach of nearly everyone who's not a tech giant. Academics and smaller companies have little choice but to get their chips from Nvidia and rent computing power from Amazon, Microsoft, or Google, and once companies are on those platforms, they are often locked in. AI start-ups frequently complain that once they start building their services on Microsoft's or Amazon's cloud services, it's difficult to switch away. Those same companies also have rare access to the thousands of GPUs needed to build anything like ChatGPT. Getting those chips, which can cost around $40,000 each, has been like trying to buy concert tickets for a sell-out tour. Nvidia, the world's main supplier of GPUs, has benefited handsomely from the demand. In May 2023, it became the latest technology company, after Google, Microsoft, Amazon, Meta, and Apple, to reach a market capitalization of $1 trillion. The biggest companies in the world by a huge margin were building tech and AI. But rather than create a thriving market for innovative new companies, the AI boom was helping these firms consolidate their power. Having strengthened their grip on infrastructure, talent, data, computing power, and profits, there is no question that they alone will control our AI future.

The AGI dreamers helped make that happen. In June 2023, Microsoft's chief financial officer Amy Hood told investors that AI-powered services from OpenAI would contribute at least $10 billion to its revenue. She called it "the fastest growing $10 billion business in our history."

Would it have been better for DeepMind and OpenAI to stay independent and assign boards of trustees to govern the direction of AI? That would have come with its own risks, as Sam Altman would later discover. Suleyman, who fought hard to extricate his company from DeepMind, has argued in interviews that large companies can be more trustworthy than smaller ones. They are, after all, publicly accountable to their shareholders and staff. But the world's biggest technology firms also have a deeper obligation to their shareholders, one that they cannot escape. They must grow their earnings each quarter. As their profits plateau or decline, so too will their stock; when the stock falls, a company can't raise capital, and executives and employees grumble or flee. The terrifying specter of decline beckons. "These entities *have* to grow," a former Microsoft executive says. "AI is the answer."

Hassabis argues that DeepMind has become a more sensible business. "It's got to the level of maturity now where I think we can improve the lives of billions of people," he says when asked about whether he thinks AGI would still need the kind of ethics board he once pushed for. “Google is an amazing place.”

Altman insists that despite pivoting OpenAI to a for-profit company, aligning with Microsoft, and sparking an AI arms race, its principles of building beneficial AI haven’t changed. And he has little choice but to keep launching AI tools to the public. “Deploying is critical to our mission,” he says. How else will OpenAI learn, and bring people useful tools like ChatGPT? That “requires putting the technology in people’s hands.”

At the current speed of this race, it’s impossible to predict what will happen in the months and years after these words were written in March 2024. But those future events will be rooted in the designs of a handful of people and the systemic forces they operated in. When the question arises of whether we can trust Sam Altman and Demis Hassabis, along with Microsoft and Google, to build our AI future, the answer is that we have little choice in the matter. Both men hitched their innovations onto two of the world’s largest companies, whose network effects were almost impossible to avoid in everyday life, and in so doing, they joined a long history of innovators who tweaked their ideals to stay in a race and build power. The result is some of the most transformative technology we have ever seen. Now to find out the price.

ACKNOWLEDGMENTS

This book could never have happened without the support and encouragement of a small army of people. About a month after ChatGPT was released to the world, I sent an idea over to my agent, David Fugate, about how two men dreamed of building superintelligent machines, then became rivals, and then proxies for a battle between Big Tech companies. David's answer, "YES!" propelled me over the following year to pull it together. Janhoi McGregor also gave me the extra kick I needed to embark on a major project, as usual over a cheeky Nando's.

I'm grateful to Claire Cheek, Sara Beth Haring, Elisa Rivlin, and all the wonderful folks at St. Martin's Press, and my editor Peter Wolverton, who provided the level-headed guidance I needed to keep things focused on how a transformative invention is being steered by a handful of companies. Pete saved me (and you, dear reader) from going down several extraneous rabbit holes around transhumanism and the effective altruism movement in this story.

Since we're on that subject, I must thank Dr. Emile Torres for brilliantly tracing the quest to build artificial general intelligence back to darker roots in eugenics and for helping me understand the flipside of chasing human perfection through machines. David Edmonds did much the same for me on the topic of long termism, Toby Ord on effective altruism, and Mike Levine on the AI alignment efforts of organizations like Open Philanthropy. Mike and I don't see eye to eye on some of these topics, but I'm very grateful to him for his patience and for taking the time to explain his arguments to me. Brian Evergreen in Seattle also helped me better understand the AI ethics conundrum that was happening inside Microsoft.

Special thanks to Meredith Whittaker and Reolof Botha, two people from very different parts of the technology world—Meredith is the president of the Signal Foundation and Reolof a partner at venture capital firm Sequoia Capital—but both helped crystalize in my mind that the growing dominance of a few technology firms in AI was becoming a problem for both society and businesses.

I've mentioned anonymous sources in my endnotes, but I must thank the many individuals once again who work for tech and AI firms, or former employees of OpenAI and DeepMind, including senior executives, who spent hours sharing their experiences and, on occasion, deep unease about the stranglehold that tech giants have acquired on the field of AI and its new move-fast-break-things dynamics. I hope this book has done justice to their concerns and all the time they shared with me.

My editors at *Bloomberg Opinion* have been incredibly supportive of this project, and I'm deeply grateful to Tim O'Brien and Nicole Torres for their enthusiasm and immediate willingness to let me take the time I needed to write a book that complemented much of what I'd been harping on about in columns for the previous year. Many thanks also to all my fellow writers at *Bloomberg Opinion* for their kind words and encouragement, and for making the job of tech columnist far more fun than it should be, including Dave Lee, Lara Williams, Lionel Laurent, Andrea Felstead, Therese Raphael, Matthew Brooker, Howard Chua-Eoan, Chris Hughes, Chris Bryant, Marcus Ashworth, Marc Champion, James Hertling, Joi Preciphs, Mark Gilbert, Elaine He, Tim Culpan, and also to Javier Blas for his advice on embarking on a book project.

I'm particularly grateful to my colleagues Tim O'Brien, Nicole Torres, Paul Davies, and Adrian Wooldridge for their sharp and helpful feedback on early versions of the manuscript, which greatly helped me understand the view of other readers outside of my writing bubble. My other early readers not only showed me where I needed to polish things up or explain things better but offered frequent moral support as great friends throughout the writing process. Many thanks for that to Miriam Zaccarelli, Victor Zaccarelli, Carley Sime, and Kristin Peterson.

Thank you to my neighbor and friend Catalina "Katina" Montesinos for allowing me to work in the quiet solitude of her home while I was on book leave. Katina is someone for whom AI can make a substantial and positive difference. A former painter who went blind at the age of forty, she became a successful sculptor and embraced all the ways technology could help her "see" the world, keenly exploiting digital assistants like Apple's Siri. At age eighty, she listened in quiet astonishment when I pointed my phone at a sculpture on her coffee table and got ChatGPT to generate a detailed analysis of its colors, shape, and possible artistic influences. It was, she said, "absolutely extraordinary." I hope that is the direction our world takes AI in the future, filling vacancies of information rather than displacing the work and creativity of humans.

Finally, this book could not have happened at all without the support of my family, including my father, Philip Withers, who has been nothing but a fount of praise for even my smallest accomplishments since I was a kid. Thank you to Cara and Wesley for filling the house with laughter and to Isla for checking in on the manuscript and cheering me on. I marvel every day at how my husband, Mani, holds us all together; he has been a constant source of strength and patience, and for that I am especially thankful.

SOURCES

A Note on Sources

I'm grateful for the extraordinary work of reporters and authors of numerous articles from websites, newspapers, magazines, research papers, podcasts, and books who have allowed me to augment this story with so many secondary sources, which I list here.

Throughout the book, where people are quoted in the present tense and cited with words such as "says," "remembers," or "recalls," these are direct interviews that I conducted with those individuals, and they include Demis Hassabis and Sam Altman. Many others whom I spoke to are referred to as former employees or individuals familiar with the matter, and who remain anonymous for reasons often related to the risk of damaging repercussions for speaking out. I am especially grateful to these people for trusting me with their insights.

There are more interviews that I could not include in the story due to constraints on space, but that were valuable in giving me context on the lives of Sam Altman and Demis Hassabis and their work and the field of AI, as well as experts who helped me understand and translate the workings of machine-learning systems, neural networks, diffusion systems, and transformers into accessible language.

My conversations with hundreds of industry experts, entrepreneurs, venture capitalists, employees and former employees of tech companies, both for this book and over the last few years of reporting about the new AI boom for *Bloomberg Opinion*, *Wall Street Journal,* and *Forbes* also greatly informed my research.

I exploited my love of running to listen to countless hours of podcast interviews with Sam Altman, Demis Hassabis, Ilya Sutskever, Greg Brockman, and many other individuals who were involved in the creation of OpenAI and DeepMind, or who witnessed the evolution of AI from scientific backwater to booming business, to help piece together many details of the narrative. That often meant frequent stops to type things into my phone, but it was worth it.

Chapter 1: High School Hero

Altman, Sam. "Machine Intelligence, Part 1." blog.samaltman.com, February 25, 2015.

Cannon, Craig. "Sam Altman." *Y Combinator* (podcast), November 8, 2018.

"First Look: Loopt Provides More Incentives to Try Location-based Services with Loopt Star." Robert Scoble's YouTube channel, June 1, 2010.

Friend, Tad. "Sam Altman's Manifest Destiny." *New Yorker*, October 3, 2016.

Graham, Paul. "How to Start a Startup." paulgraham.com, March 2005.

Graham, Paul. "How Y Combinator Started." paulgraham.com, March 2012.

Graham, Paul. "The Word 'Hacker.'" paulgraham.com, April 2024.

"How Tesla Became the Elon Musk Co." *Land of the Giants* (podcast), Vox Media, August 2, 2023.

Internet Archive for the now defunct website, http://www.loopt.com/.

Lessin, Jessica. "This Is How Sam Altman Works the Press and Congress. I Know from Experience." *The Information*, June 7, 2023.

Mitchell, Melanie. *Artificial Intelligence: A Guide for Thinking Humans*. New York: Pelican, 2020.

Wagstaff, Keith. "The Good Ol' Days of AOL Chat Rooms." *CNN*, July 6, 2012.

Weil, Elizabeth. "Sam Altman Is the Oppenheimer of Our Age." *New York Magazine*, September 25, 2023.

Wired. "Sebastian Thrun & Sam Altman Talk Flying Vehicles and Artificial Intelligence." Video of *Wired* conference panel, October 16, 2018.

"WWDC 2008 News: Loopt Shows Off New App for the iPhone." CNET's YouTube channel, June 10, 2008.

Chapter 2: Winning, Winning, Winning

"A.I. Could Solve Some of Humanity's Hardest Problems. It Already Has." *The Ezra Klein Show*, July 11, 2023.

Burton-Hill, Clemency. "The Superhero of Artificial Intelligence." *The Guardian*, February 16, 2016.

"Demis Hassabis." *Desert Island Discs* (podcast), BBC Radio 4, May 21, 2017.

"Demis Hassabis, Ph.D." *What It Takes* (podcast), American Academy of Achievement, April 23, 2018.

"Genius Entrepreneur." *The Bridge*, Queens College Cambridge Magazine, September 2014.

"Google DeepMind's Demis Hassabis." *The Bottom Line* (podcast), BBC Radio 4, October 16, 2023.

Hassabis, Demis. *The Elixir Diaries*, columns in *Edge* magazine, also available at https://archive.kontek.net/, 1998–2000.

Parker, Sam. "Republic: The Revolution Review." *GameSpot*, September 2, 2003.

Pearce, Jacqui. "Getting to Know You," a Q&A with Angela Hassabis, *HBC Accord*, (Hendon Baptist Church newsletter), October 2018.

"Republic." *Edge* magazine, November 1999.

Weinberg, Steven. *Dreams of a Final Theory*. New York: Vintage, 1994.

Chapter 3: Save the Humans

Altman, Sam. "Hard Tech Is Back." blog.samaltman.com, March 11, 2016.

Altman, Sam. "Startup Advice." blog.samaltman.com, June 3, 2013.

Altman, Sam. "YC and Hard Tech Startups." ycombinator.com, date not provided.

Cannon, Craig. "Sam Altman." *Y Combinator* (podcast), November 8, 2018.

Chafkin, Max. "Y Combinator President Sam Altman Is Dreaming Big." *Fast Company*, April 16, 2015.

Clifford, Catherine. "Nuclear Fusion Start-Up Helion Scores $375 Million Investment from Open AI CEO Sam Altman." *CNBC*, November 5, 2021.

Dwoskin, Elizabeth, Marc Fisher, and Nitasha Tiku. "'King of the Cannibals': How Sam Altman Took Over Silicon Valley." *Washington Post*, December 23, 2023.

Friend, Tad. "Sam Altman's Manifest Destiny." *New Yorker*, October 3, 2016.

Graham, Paul. "Five Founders." paulgraham.com, April 2019.

Graham, Paul. "How to Start a Startup." paulgraham.com, March 2005.

“How and Why to Start a Startup—Sam Altman & Dustin Moskovitz—Stanford CS183F: Startup School.” Stanford Online’s YouTube channel, April 5, 2017.

“Paul Graham on Why Sam Altman Took Over as President of Y Combinator in 2014.” This Week in Startups Clips’s YouTube channel, March 19, 2019.

Regalado, Antonio. “A Startup Is Pitching a Mind-Uploading Service that Is ‘100 Percent Fatal.’” *MIT Technology Review*, March 13, 2018.

“Sam Altman—Leading with Crippling Anxiety, Discovering Meditation, and Building Intelligence with Self-Awareness.” *The Art of Accomplishment* (podcast), January 15, 2022.

Stiegler, Marc. *The Gentle Seduction*. New York: Baen, 1990.

Chapter 4: A Better Brain

Bostrom, Nick. *Superintelligence: Paths, Dangers, Strategies*. Oxford: Oxford University Press, 2014.

Cutright, Keisha M., and Mustafa Karataş. “Thinking about God Increases Acceptance of Artificial Intelligence in Decision-Making.” *Proceedings of the National Academy of Sciences (PNAS)* 120, no. 33 (2023): e2218961120–e2218961120.

“Demis Hassabis, Ph.D.” *What It Takes* (podcast), American Academy of Achievement, April 23, 2018.

Dowd, Maureen. “Elon Musk’s Billion-Dollar Crusade to Stop the A.I. Apocalypse.” *Vanity Fair*, March 26, 2017.

Goertzel, Ben. *AGI Revolution: An Inside View of the Rise of Artificial General Intelligence*. Middletown, DE: Humanity+ Press, 2016.

Hassabis, Demis. “The Neural Processes Underpinning Episodic Memory.” PhD thesis, University College London, February 2009.

Homer-Dixon, Thomas. *The Ingenuity Gap*. New York: Knopf Doubleday, 2000.

Kurzweil, Ray. *The Age of Spiritual Machines*. New York: Penguin, 2000.

McCarthy, John, Marvin L. Minsky, Nathaniel Rochester, and Claude E. Shannon. “A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence: August 31, 1955.” *The AI Magazine* 27, no. 4 (2006): 12–14. Copies of the original typescript are housed in the archives at Dartmouth College and Stanford University. A reproduction of the proposal can be found here: https://ojs.aaai.org/aimagazine/index.php/aimagazine/article/view/1904.

Penrose, Roger. *Shadows of the Mind: A Search for the Missing Science of Consciousness*. London: Vintage, 1995.

Syed, Matthew. “Demis Hassabis Interview: The Kid from the Comp Who Founded DeepMind and Cracked a Mighty Riddle of Science.” *The Sunday Times*, December 5, 2020.

“A Systems Neuroscience Approach to Building AGI—Demis Hassabis, Singularity Summit 2010.” Google DeepMind’s YouTube channel, March 7, 2018.

Thiel, Peter, with Blake Masters. *Zero to One: Notes on Start Ups, or How to Build the Future*. London: Virgin Books, 2015.

Chapter 5: For Utopia, for Money

“Andrew Ng: Deep Learning, Education, and Real-World AI.” *Lex Fridman Podcast* (podcast), February 20, 2020.

“Bill Gates, Sergey Brin, and Larry Page: Tech Titans.” *What It Takes* (podcast), American Academy of Achievement, achievement.org, January 13, 2020.

Copeland, Rob. “Google Management Shuffle Points to Retreat from Alphabet Experiment.” *Wall Street Journal*, December 5, 2019.

Hodson, Hal. “DeepMind and Google: The Battle to Control Artificial Intelligence.” *Economist*, March 1, 2019.

Huxley, Julian. “Transhumanism.” *Journal of Humanistic Psychology*, January 1968.

Markram, Henry. “A Brain in a Supercomputer.” TED Global, July 2009.

Metz, Cade. *Genius Makers: The Mavericks Who Brought A.I. to Google, Facebook, and the World*. New York: Random House Business, 2021.

“Peter Thiel Says America Has Bigger Problems Than Wokeness.” *Honestly with Bari Weiss* (podcast), May 3, 2023.

Suleyman, Mustafa, with Michael Bhaskar. *The Coming Wave*. New York: Crown, 2023.

Chapter 6: The Mission

Albergotti, Reed. “The Secret History of Elon Musk, Sam Altman, and OpenAI.” *Semafor*, March 24, 2023.

Birhane, Abeba, Pratyusha Kalluri, Dallas Card, William Agnew, Ravit Dotan, and Michelle Bao. “The Values Encoded in Machine Learning Research.” *FAccT Conference ’22: Proceedings of the 2022 ACM Conference on Fairness*, *Accountability, and Transparenc*y (June 2022): 173–84.

Brockman, Greg. “My Path to OpenAI.” blog.gregbrockman.com, May 3, 2016.

Conn, Ariel. “Concrete Problems in AI Safety with Dario Amodei and Seth Baum.” *Future of Life Institute* (podcast), August 31, 2016.

Dowd, Maureen. “Elon Musk’s Billion-Dollar Crusade to Stop the A.I. Apocalypse.” *Vanity Fair*, March 26, 2017.

Elon Musk vs Sam Altman [2024] CGC-24-612746.

Friend, Tad. “Sam Altman’s Manifest Destiny.” *New Yorker*, October 3, 2016.

Galef, Jesse. “Elon Musk Donates $10M to Our Research Program.” futureoflife.org, January 22, 2015.

“Greg Brockman: OpenAI and AGI.” *Lex Fridman Podcast* (podcast), April 3, 2019.

Harris, Mark. “Elon Musk Used to Say He Put $100M in OpenAI, but Now It’s $50M: Here Are the Receipts.” *TechCrunch*, May 18, 2023.

Metz, Cade. “Ego, Fear and Money: How the A.I. Fuse Was Lit.” *New York Times*, December 3, 2023.

Metz, Cade. “Inside OpenAI, Elon Musk’s Wild Plan to Set Artificial Intelligence Free.” *Wired*, April 27, 2016.

Vance, Ashlee. *Elon Musk: How the Billionaire CEO of SpaceX and Tesla Is Shaping Our Future*. London: Virgin Books, 2015.

Chapter 7: Playing Games

Aron, Jacob. “How to Build the Global Mathematics Brain.” *New Scientist*, May 4, 2011.

Byford, Sam. “Google’s AlphaGo AI Defeats World Go Number One Ke Jie.” *The Verge*, May 23, 2017.

Gallagher, Ryan. “Google’s Secret China Project ‘Effectively Ended’ after Internal Confrontation.” *The Intercept*, December 17, 2018.

Gallagher, Ryan. “Private Meeting Contradicts Google’s Official Story on China.” *The Intercept*, October 9, 2018.

“Has Anyone Actually Tried to Convince Terry Tao or Other Top Mathematicians to Work on Alignment?” www.lesswrong.com, June 8, 2022.

“How to Play.” British Go Association, updated October 26, 2017, https://www.britgo.org/intro/intro2.html.

Metz, Cade. *Genius Makers: The Mavericks Who Brought A.I. to Google, Facebook, and the World*. New York: Random House Business, 2021.

Metz, Cade. “Google Is Already Late to China’s AI Revolution.” *Wired*, June 2, 2017.

Rogin, Josh. “Eric Schmidt: The Great Firewall of China Will Fall.” *Foreign Policy*, July 9, 2012.

Suleyman, Mustafa, with Michael Bhaskar. *The Coming Wave*. New York: Crown, 2023.

Temperton, James. “DeepMind’s New AI Ethics Unit Is the Company’s Next Big Move.” *Wired*, October 4, 2017.

Yang, Yuan. “Google’s AlphaGo Is World’s Best Go Player.” *Financial Times*, May 25, 2017.

Chapter 8: Everything Is Awesome

Angwin, Julia, Jeff Larson, Surya Mattu, and Lauren Kirchner. “Machine Bias.” *ProPublica*, May 23, 2016.

Buolamwini, Joy, and Timnit Gebru. “Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification.” *Proceedings of Machine Learning Research* 81 (2018): 1–15.

Dastin, Jeffrey. “Amazon Scraps Secret AI Recruiting Tool That Showed Bias against Women.” *Reuters*, October 10, 2018.

Devlin, Hannah, and Alex Hern. “Why Are There So Few Women in Tech? The Truth behind the Google Memo.” *The Guardian*, August 8, 2017.

Gebru, Timnit, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. “Datasheets for Datasets.” *Communications of the ACM* 64, no. 12 (2021): 86–92.

Grant, Nico, and Kashmir Hill. “Google’s Photo App Still Can’t Find Gorillas. And Neither Can Apple’s.” *New York Times*, May 22, 2023.

Harris, Josh. “There Was All Sorts of Toxic Behaviour’: Timnit Gebru on Her Sacking by Google, AI’s Dangers and Big Tech’s Biases.” *The Guardian*, May 22, 2023.

Horwitz, Jeff. “The Facebook Files.” *Wall Street Journal*, October 1, 2021.

Payton, L’Oreal Thompson. “Americans Check Their Phones 144 Times a Day. Here’s How to Cut Back.” *Fortune*, July 19, 2023.

Simonite, Tom. “What Really Happened When Google Ousted Timnit Gebru.” *Wired*, June 8, 2021.

“The Social Atrocity: Meta and the Right to Remedy for the Rohingya.” Amnesty International report, September 29, 2022.

Wakabayashi, Daisuke, and Katie Benner. “How Google Protected Andy Rubin, the ‘Father of Android.’” *New York Times*, October 25, 2018.

Chapter 9: The Goliath Paradox

de Vynck, Gerrit. “Google’s Cloud Unit Won’t Sell a Type of Facial Recognition Tech.” *Bloomberg*, December 13, 2018.

“Google Duplex: A.I. Assistant Calls Local Businesses to Make Appointments.” Jeff Grubb’s Game Mess’s YouTube channel, May 8, 2018.

Kruppa, Miles, and Sam Schechner. “How Google Became Cautious of AI and Gave Microsoft an Opening.” *Wall Street Journal*, March 7, 2023.

Love, Julia. “Google Says Over Half of Generative AI Startups Use Its Cloud.” *Bloomberg*, August 29, 2023.

Nylen, Leah. “Google Paid $26 Billion to Be Default Search Engine in 2021.” *Bloomberg*, October 17, 2021.

Uszkoreit, Jakob. “Transformer: A Novel Neural Network Architecture for Language Understanding.” blog.research.google, August 31, 2017.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. “Attention Is All You Need.” *Advances in Neural Information Processing Systems* 30 (2017).

Chapter 10: Size Matters

Brockman, Greg (@gdb). “Held our civil ceremony in the @OpenAI office last week. Officiated by @ilyasut, with the robot hand serving as ring bearer. Wedding planning to commence soon.” Twitter, November 12, 2019, 9:39 a.m. https://twitter.com/gdb/status/1194293590979014657?lang=en.

Brockman, Greg. “Microsoft Invests in and Partners with OpenAI to Support Us Building Beneficial AGI.” www.openai.com, July 22, 2019.

“Greg Brockman: OpenAI and AGI.” *Lex Fridman Podcast* (podcast), April 3, 2019.

Hao, Karen. “The Messy, Secretive Reality behind OpenAI’s Bid to Save the World.” *MIT Technology Review*, February 17, 2020.

Hao, Karen, and Charlie Warzel. “Inside the Chaos at OpenAI.” *The Atlantic*, November 19, 2023.

Jin, Berber, and Keach Hagey. “The Contradictions of Sam Altman, AI Crusader.” *Wall Street Journal*, March 31, 2023.

Kraft, Amy. “Microsoft Shuts Down AI Chatbot after It Turned into a Nazi.” *CBS News*, March 25, 2016.

Levy, Steven. “What OpenAI Really Wants.” *Wired*, September 5, 2023.

Metz, Cade. “A.I. Researchers Are Making More Than $1 Million, Even at a Nonprofit.” *New York Times*, April 19, 2018.

Metz, Cade. “The ChatGPT King Isn’t Worried, but He Knows You Might Be.” *New York Times*, March 31, 2023.

“OpenAI Charter.” www.openai.com/charter, April 9, 2018.

Radford, Alec, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. “Improving Language Understanding by Generative Pre-Training.” www.openai.com, June 11, 2018.

Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. “Language Models Are Unsupervised Multitask Learners.” www .openai.com, February 14, 2019.

Chapter 11: Bound to Big Tech

Ahmed, Nur, Muntasir Wahed, and Neil C. Thompson. “The Growing Influence of Industry in AI Research.” *Science*, March 2, 2023.

Amodei, Dario, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. “Concrete Problems in AI Safety.” www.arxiv.org, July 25, 2016.

Copeland, Rob. “Google Management Shuffle Points to Retreat from Alphabet Experiment.” *Wall Street Journal*, December 5, 2019.

Coulter, Martin, and Hugh Langley. “DeepMind’s Cofounder Was Placed on Leave after Employees Complained about Bullying and Humiliation for Years. Then Google Made Him a VP.” *Business Insider*, August 7, 2021.

Friend, Tad. “Sam Altman’s Manifest Destiny.” *New Yorker*, October 3, 2016.

Hodson, Hal. “Revealed: Google AI Has Access to Huge Haul of NHS Patient Data.” *New Scientist*, April 29, 2016.

Ludlow, Edward, Matt Day, and Dina Bass. “Amazon to Invest Up to $4 Billion in AI Startup Anthropic.” *Bloomberg*, September 25, 2023.

Piper, Kelsey. “Exclusive: Google Cancels AI Ethics Board in Response to Outcry.” *Vox*, April 4, 2019.

Primack, Dan. “Google Is Investing $2 Billion into Anthropic, a Rival to OpenAI.” *Axios*, October 30, 2023.

Waters, Richard. “DeepMind Co-founder Leaves Google for Venture Capital Firm.” *Financial Times*, January 21, 2022.

Chapter 12: Myth Busters

Abid, Abubakar, Maheen Farooqi, and James Zou. “Large Language Models Associate Muslims with Violence.” *Nature Machine Intelligence* 3 (2021): 461–63.

Barrett, Paul, Justin Hendrix, and Grant Sims. “How Tech Platforms Fuel U.S. Political Polarization and What Government Can Do about It.” www.brookings.edu, September 27, 2021.

Bender, Emily, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. “On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?” *FAccTConference ’21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (March 2021): 610–23. https://dl.acm.org/doi/10.1145/3442188.3445922.

Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. “Language Models Are Few-Shot Learners.” www.openai.com, July 22, 2020.

Gehman, Samuel, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. “RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models.” *ACL Anthology*. Findings of the Association for Computational Linguistics: EMNLP 2020, November 2020.

Hornigold, Thomas. “This Chatbot Has Over 660 Million Users—and It Wants to Be Their Best Friend.” *Singularity Hub*, July 14, 2019.

Jin, Berber, and Miles Kruppa. “Microsoft to Deepen OpenAI Partnership, Invest Billions in ChatGPT Creator.” *Wall Street Journal*, January 23, 2023.

Lecher, Colin. “The Artificial Intelligence Field Is Too White and Too Male, Researchers Say.” *The Verge*, April 17, 2019.

Lemoine, Blake. “I Worked on Google’s AI. My Fears Are Coming True.” *Newsweek*, February 27, 2023.

Lodewick, Colin. “Google’s Suspended AI Engineer Corrects the Record: He Didn’t Hire an Attorney for the ‘Sentient’ Chatbot, He Just Made Introductions—the Bot Hired the Lawyer.” *Fortune*, June 23, 2022.

Luccioni, Alexandra, and Joseph Viviano. “What’s in the Box? An Analysis of Undesirable Content in the Common Crawl Corpus.” *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*. Volume 2: Short Papers (2021): 182–89.

Muller, Britney. “BERT 101: State of the Art NLP Model Explained.” www.huggingface.co, March 2, 2022.

Newton, Casey. “The Withering Email That Got an Ethical AI Researcher Fired at Google.” *Platformer*, December 3, 2020.

Nicholson, Jenny. “The Gender Bias Inside GPT-3.” www.medium.com, March 8, 2022.

Perrigo, Billy. “Exclusive: OpenAI Used Kenyan Workers on Less Than $2 Per Hour to Make ChatGPT Less Toxic.” *Time*, January 18, 2023.

Silverman, Craig, Craig Timberg, Jeff Kao, and Jeremy B. Merrill. “Facebook Hosted Surge of Misinformation and Insurrection Threats in Months Leading Up to Jan. 6 Attack, Records Show.” *ProPublica* and *Washington Post*, January 4, 2022.

Simonite, Tom. “What Really Happened When Google Ousted Timnit Gebru.” *Wired*, June 8, 2021.

Tiku, Nitasha. “The Google Engineer Who Thinks the Company’s AI Has Come to Life.” *Washington Post*, June 11, 2022.

Venkit, Pranav Narayanan, Mukund Srinath, and Shomir Wilson. “A Study of Implicit Language Model Bias against People with Disabilities.” *Proceedings of the 29th International Conference on Computational Linguistics* (2022): 1324–32.

Wendler, Chris, Veniamin Veselovsky, Giovanni Monea, and Robert West. “Do Llamas Work in English? On the Latent Language of Multilingual Transformers.” www.arxiv.org, February 16, 2024.

Chapter 13: Hello, ChatGPT

“AlphaFold: The Making of a Scientific Breakthrough.” Google DeepMind’s YouTube channel, November 30, 2020.

Andersen, Ross. “Does Sam Altman Know What He’s Creating?” *The Atlantic*, July 24, 2023.

Grant, Nico. “Google Calls in Help from Larry Page and Sergey Brin for A.I. Fight.” *New York Times*, January 20, 2023.

Grant, Nico, and Cade Metz. “A New Chat Bot Is a ‘Code Red’ for Google’s Search Business.” *New York Times*, December 21, 2022.

Hao, Karen, and Charlie Warzel. “Inside the Chaos at OpenAI.” *The Atlantic*, November 19, 2023.

Heikkilä, Melissa. “This Artist Is Dominating AI-generated Art. And He’s Not Happy About It.” *MIT Technology Review*, September 16, 2022.

“Introducing ChatGPT.” www.openai.com, November 30, 2022.

Johnson, Khari. “DALL-E 2 Creates Incredible Images—and Biased Ones You Don’t See.” *Wired*, May 5, 2022.

McLaughlin, Kevin, and Aaron Holmes. “How Microsoft’s Stumbles Led to Its OpenAI Alliance.” *The Information*, January 23, 2023.

Merritt, Rick. “AI Opener: OpenAI’s Sutskever in Conversation with Jensen Huang.” www.blogs.nvidia.com, March 22, 2023.

“Microsoft CTO Kevin Scott on AI Copilots, Disagreeing with OpenAI, and Sydney Making a Comeback.” *Decoder with Nilay Patel* (podcast), May 23, 2023.

Patel, Nilay. “Microsoft Thinks AI Can Beat Google at Search—CEO Satya Nadella Explains Why.” *The Verge*, February 8, 2023.

Pichai, Sundar. “Google DeepMind: Bringing Together Two World-Class AI Teams.” www.blog.google, April 20, 2023.

Rawat, Deeksha. “Unravelling the Dynamics of Diffusion Model: From Early Concept to Cutting-Edge Applications.” www.medium.com, August 5, 2023.

Roose, Kevin. “Bing’s A.I. Chat: ‘I Want to Be Alive.’” *New York Times*, February 16, 2023.

“Sam Altman on the A.I. Revolution, Trillionaires and the Future of Political Power.” *The Ezra Klein Show* (podcast), June 11, 2021.

Weise, Karen, Cade Metz, Nico Grant, and Mike Isaac. “Inside the A.I. Arms Race That Changed Silicon Valley Forever.” *New York Times*, December 5, 2023.

Chapter 14: A Vague Sense of Doom

Details about Open Philanthropy’s disclosure of its executive director being married to someone who worked at OpenAI comes from www.openphilanthropy.org/grants/openai-general-support/.

Details of investments by FTX founders into Anthropic come from Pitchbook, a market research firm.

Details on Open Philanthropy’s grants and funding come from www.openphilanthropy.org/grants/.

Texts between William MacAskill and Elon Musk are sourced from court filings that were released as part of a pretrial discovery process in a legal battle between Musk and Twitter, dated September 28, 2022.

Anderson, Mark. “Advice for CEOs Under Pressure from the Board to Use Generative AI.” *Fast Company*, October 31, 2023.

Berg, Andrew, Christ Papageorgiou, and Maryam Vaziri. “Technology’s Bifurcated Bite.” *F&D Magazine*, International Monetary Fund, December 2023.

Bordelon, Brendan. “How a Billionaire-Backed Network of AI Advisers Took Over Washington.” *Politico*, February 23, 2024.

“EU AI Act: First Regulation on Artificial Intelligence.” www.europarl.europa.eu, June 8, 2023.

Gross, Nicole. “What ChatGPT Tells Us about Gender: A Cautionary Tale about Performativity and Gender Biases in AI.” *Social Sciences*, August 1, 2023.

Johnson, Simon, and Daron Acemoglu. *Power and Progress: Our Thousand-Year Struggle Over Technology and Prosperity*. New York: Basic Books, 2023.

Lewis, Gideon. “The Reluctant Prophet of Effective Altruism.” *New Yorker*, August 8, 2022.

Lewis, Michael. *Going Infinite*. New York: Penguin, 2023.

MacAskill, William. *What We Owe the Future*. London: Oneworld, 2022.

Metz, Cade. “The ChatGPT King Isn’t Worried, but He Knows You Might Be.” *New York Times*, March 31, 2023.

Metz, Cade. “‘The Godfather of A.I.’ Leaves Google and Warns of Danger Ahead.” *New York Times*, May 1, 2023.

Millar, George. “The Magical Number Seven, Plus or Minus Two.” *Psychological Review*, 1956.

Milmo, Dan, and Alex Hern. “Discrimination Is a Bigger AI Risk Than Human Extinction—EU Commissioner.” *The Guardian*, June 14, 2023.

Mollman, Steve. “A Lawyer Fired after Citing ChatGPT-Generated Fake Cases Is Sticking with AI Tools.” *Fortune*, November 17, 2023.

Moss, Sebastian. “How Microsoft Wins.” www.datacenterdynamics.com, November 24, 2023.

O’Brien, Sara Ashley. “Bumble CEO Whitney Wolfe Herd Steps Down.” *Wall Street Journal*, November 6, 2023.

“Pause Giant AI Experiments: An Open Letter.” Future of Life Institute, www .futureoflife.org, March 22, 2023.

Perrigo, Billy. “OpenAI Could Quit Europe Over New AI Rules, CEO Sam Altman Warns.” *Time*, May 25, 2023.

Piantadosi, Steven (@spiantado). “Yes, ChatGPT is amazing and impressive. No, @OpenAI has not come close to addressing the problem of bias. Filters appear to be bypassed with simple tricks, and superficially masked.” Twitter, December 4, 2022, 10:55 a.m. https://twitter.com/spiantado/status/1599462375887114240?lang=en.

Piper, Kelsey. “Sam Bankman-Fried Tries to Explain Himself.” *Vox*, November 16, 2022.

“Rishi Sunak & Elon Musk: Talk AI, Tech & the Future.” Rish Sunak’s YouTube channel, November 3, 2023.

“Romney Leads Senate Hearing on Addressing Potential Threats Posed by AI, Quantum Computing, and Other Emerging Technology.” www.romney.senate.gov, September 19, 2023.

Roose, Kevin. “Inside the White-Hot Center of A.I. Doomerism.” *New York Times*, July 11, 2023.

“Sam Altman: ‘I Trust Answers Generated by ChatGPT Least than Anybody Else on Earth.’” Business Today’s YouTube channel, June 8, 2023.

Singer, Peter. *The Life You Can Save*. New York: Random House, 2010.

“Statement on AI Risk.” Center for AI Safety, www.safe.ai, May 2023.

Vallance, Chris. “Artificial Intelligence Could Lead to Extinction, Experts Warn.” *BBC News*, May 30, 2023.

Vincent, James. “OpenAI Sued for Defamation after ChatGPT Fabricates Legal Accusations against Radio Host.” *The Verge*, June 9, 2023.

Weprin, Alex. “Jeffrey Katzenberg: AI Will Drastically Cut Number of Workers It Takes to Make Animated Movies.” *Hollywood Reporter*, November 9, 2023.

Yudkowsky, Eliezer. “Pausing AI Developments Isn’t Enough. We Need to Shut It All Down.” *Time*, March 29, 2023.

Chapter 15: Checkmate

“The Capabilities of Multimodal AI|Gemini Demo.” Google’s YouTube channel, December 6, 2023.

Dastin, Jeffrey, Krystal Hu, and Paresh Dave. “Exclusive: ChatGPT Owner OpenAI Projects $1 Billion in Revenue by 2024.” *Reuters*, December 15, 2022.

Gurman, Mark. “Apple’s iPhone Design Chief Enlisted by Jony Ive, Sam Altman to Work on AI Devices.” *Bloomberg*, December 26, 2023.

Hagey, Keach, Deepa Seetharaman, and Berber Jin. “Behind the Scenes of Sam Altman’s Showdown at OpenAI.” *Wall Street Journal*, November 22, 2023.

Hawkins, Mackenzie, Edward Ludlow, Gillian Tan, and Dina Bass. “OpenAI’s Sam Altman Seeks US Blessing to Raise Billions for AI Chips.” *Bloomberg*, February 16, 2024.

Heath, Alex. “Mark Zuckerberg’s New Goal Is Creating Artificial General Intelligence.” *The Verge*, January 18, 2024.

Imbrie, Andrew, Owen Daniels, and Helen Toner. “Decoding Intentions: Artificial Intelligence and Costly Signals.” Center for Security and Emerging Technology, October 2023.

Metz, Cade, Tripp Mickle, and Mike Isaac. “Before Altman’s Ouster, OpenAI’s Board Was Divided and Feuding.” *New York Times*, November 21, 2023.

Roose, Kevin. “Inside the White-Hot Center of A.I. Doomerism.” *New York Times*, July 11, 2023.

Sigalos, MacKenzie, and Ryan Browne. “OpenAI’s Sam Altman Says Human-level AI Is Coming but Will Change World Much Less Than We Think.” *CNBC*, January 16, 2024.

Victor, Jon, and Amir Efrati. “OpenAI Made an AI Breakthrough before Altman Firing, Stoking Excitement and Concern.” *The Information*, November 22, 2023.

Walker, Bernadette. “Inside OpenAI’s Shock Firing of Sam Altman.” *Bloomberg*, November 20, 2023.

Zuckerberg, Mark. “Some Updates on Our AI Efforts.” Video posted January 18, 2024 on Facebook. https://www.facebook.com/zuck/posts/pfbid02UhntmXwNBLiV8EZHK71gAQmTx8i4vhfte9vfqjrqyGytfuW4dPQSQ5BnbzMBSPY5l.

Chapter 16: In the Shadow of Monopolies

Bommasani, Rishi, Kevin Klyman, Shayne Longpre, Sayash Kapoor, Nestor Maslej, Betty Xiong, Daniel Zhang, and Percy Liang. "The Foundation Model Transparency Index." Stanford Center for Research on Foundation Models (CRFM) and Stanford Institute for Human-Centered Artificial Intelligence (HAI), October 18, 2023.

Cheng, Michelle. "AI Girlfriend Bots Are Already Flooding OpenAI's GPT Store." *Quartz*, January 11, 2024.

Cheng, Michelle. "A Startup Founded by Former Google Employees Claims that Users Spend Two Hours a Day with Its AI Chatbots." *Quartz*, October 12, 2023.

Holmes, Aaron. "Microsoft CFO Says OpenAI and Other AI Products Will Add $10 Billion to Revenue." *The Information*, June 2023.

"Introducing the GPT Store." www.openai.com, January 10, 2024.

Leswing, Kif. "Nvidia's AI Chips Are Selling for More than $40,000 on eBay." *CNBC*, April 14, 2023.

"The Long-Term Benefit Trust." www.anthropic.com/news/the-long-term-benefit-trust, September 19, 2023.

Schiffmann, Avi (@AviSchiffmann). "I just built the world's most personal wearable AI! You can talk to Tab about anything in your life. Our computers are now our creative partners!" [demo of Tab]. Twitter, October 1, 2023, 5:12 a.m. https://twitter.com/AviSchiffmann/status/1708439854005321954?lang=en.

Vance, Ashlee. "Elon Musk's Brain Implant Startup Is Ready to Start Surgery." *Bloomberg Businessweek*, November 7, 2023.

INDEX

The index that appeared in the print version of this title does not match the pages in your e-book. Please use the search function on your e-reading device to search for terms of interest. For your reference, the terms that appear in the print index are listed below.

ALSO BY PARMY OLSON

We Are Anonymous: Inside the Hacker World of LulzSec, Anonymous, and the Global Cyber Insurgency

ABOUT THE AUTHOR

Kim Farinha

Parmy Olson is a *Bloomberg Opinion* columnist who has been covering the technology industry for more than thirteen years. A former reporter for the *Wall Street Journal* and *Forbes*, she is the author of *We Are Anonymous* and a recipient of the Palo Alto Networks Cybersecurity Canon Award. Her reporting on Facebook’s $19 billion acquisition of WhatsApp and the subsequent fallout resulted in two honorable mentions in the SABEW business journalism awards. At the *Wall Street Journal* she was the first to report on a secret effort at Google’s top AI lab to break away from the company. This book expands on the dramatic, unstoppable forces that led to that incident. You can sign up for email updates here.

First published in the United States by St. Martin's Press, an imprint of St. Martin's Publishing Group

www.stmartins.com

Cover design by Rob Grom

Cover art: Portrait of Napoleon Bonaparte by François Gérard © Peter Barritt / Alamy Stock Photo

eISBN 9781250337757

First Edition: 2024

CONTENTS