PARMY OLSON

SUPREMACY

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

St. Martin's Press New York

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.

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.