PARMY OLSON

SUPREMACY

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

St. Martin's Press New York

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.

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.