Community Reading Group Annotations *August 2025*

Empire of AI: Dreams and Nightmares in Sam Altman's Open AI
By: Karen Hao

Community for me is a term of ever shifting scale. I fluidly define my community as my closest friendships, my wider social networks, people laboring in other parts of the world creating the physical/technological objects of my daily life, and all the non-human beings and environments that share the same air, water and earth as me.

As the creation and use of AI is ever more pervasive in all of these communities, I'm looking for a deeper understanding of it and a responsible engagement in the topic that moves beyond my individual actions. Even if I personally never use AI intentionally, AI is a force in the world that has become unavoidable.

– Tracy Jarvis

Karen Hao	Empire of AI	Prol. & Ch. 1	Excerpted by Andrew M.

AI is one of the most consequential technologies of this era. In a little over a decade, it has reformed the backbone of the internet, becoming a ubiquitous mediator of digital activities. In even less time, it is now on track to rewire a great many other critical functions in society, from health care to education, from law to finance, from journalism to government. The future of AI—the shape that this technology takes—is inextricably tied to our future. The question of how to govern AI, then, is really a question about how to ensure we make our future better, not worse.

From the beginning, OpenAI had presented itself as a bold experiment in answering this question. It was founded by a group including Elon Musk and Sam Altman, with other billionaire backers like Peter Thiel, to be more than just a research lab or a company. The founders asserted a radical commitment to develop so-called artificial general intelligence, what they described as the most powerful form of AI anyone had ever seen, not for the financial gains of shareholders but for the benefit of humanity. To that end, Musk and Altman had set it up as a nonprofit and pledged $1 billion for its operation.

Merely a year and a half in, OpenAI's executives realized that the path they wanted to take in AI development would demand extraordinary amounts of money. Musk and Altman, who had until then both taken more hands-off approaches as cochairmen, each tried to install himself as CEO. Altman won out. … Altman reformulated OpenAI's legal structure. Nested within the nonprofit, he created a for-profit arm, OpenAI LP, to raise capital, commercialize products, and provide returns to investors much like any other company. Four months later, in July 2019, OpenAI announced a new $1 billion funder: software giant and cloud services provider Microsoft. **(22–24)**

Broad themes: Founder myths, cults of personality, protoge culture.

See: Matt Krisiloff, Conception. Working on stem cell generation of eggs.

Techno fatalist discourse in Silicon V.

Through my reporting, I've come to understand two things: Artificial intelligence is a technology that takes many forms. It is in fact a multitude of technologies that shape-shift and evolve, not merely based on technical merit but with the ideological drives of the people who create them and the winds of hype and commercialization. While ChatGPT and other so-called large language models or generative AI applications have now taken the limelight, they are but one manifestation of AI, a manifestation that embodies a particular and remarkably narrow view about the way the world is and the way it should be. Nothing about this form of AI coming to the fore or even existing at all was inevitable; it was the culmination of thousands of subjective choices, made by the people who had the power to be in the decision-making room. In the same way, future generations of AI technologies are not predetermined. But the question of governance returns: Who will get to shape them?

The other thing I've learned: This current manifestation of AI, and the trajectory of its development, is headed in an alarming direction. On the surface, generative AI is thrilling. ... [But under] the hood. generative AI models are monstrosities, built from consuming previously unfathomable amounts of data, labor, computing power, and natural resources. ... The exploding human and material costs are settling onto wide swaths of society, especially the most vulnerable, people I met around the world, whether workers and rural residents in the Global North or impoverished communities in the Global South, all suffering new degrees of precarity. Rarely have they seen any "trickle-down" gains of this so-called technological revolution; the benefits of generative AI mostly accrue upward.

★ Over the years, I've found only one metaphor that encapsulates the nature of what these AI power players are: empires. During the long era of European colonialism, empires seized and extracted resources that were not their own and exploited the labor of the people they subjugated to mine, cultivate, and refine those resources for the empires' enrichment. They projected racist, dehumanizing ideas of their own superiority and modernity to justify—and even entice the conquered into accepting—the invasion of sovereignty, the theft, and the subjugation. They justified their quest for power by the need to compete with other empires: In an arms race, all bets are off. All this ultimately served to entrench each empire's power and to drive its expansion and progress. In the simplest terms, empires amassed extraordinary riches across space and time, through imposing a colonial world order, at great expense to everyone else. **(25–27)**

"Empire" might be useful in rendering visible the reach of these technologies (e.g., resource extaction, political influence, etc.)

On the other hand, as a metaphor it might be too reductive.

Another important assoc is that empires are full of involuntary subjects.

Everyone else had arrived, but Elon Musk was late as usual. It was the summer of 2015, and a group of men had gathered for a private dinner at Sam Altman's invitation to discuss the future of AI and humanity.

Musk had met Altman, fourteen years his junior, a while earlier and had formed a good impression. President of the famed Silicon Valley startup accelerator Y Combinator, Altman's reputation preceded him. After starting his first company at age nineteen, he had rapidly established himself within Silicon Valley as a brilliant strategist and dealmaker with grand ambitions, even for the land of big-thinking founders.

[Just a month or so prior to the meeting] Altman emailed Musk. "Been thinking a lot about whether it's possible to stop humanity from developing AI," Altman wrote. "I think the answer is almost definitely not. If it's going to happen anyway, it seems like it would be good for someone other than Google to do it first." He proposed for Y Combinator, or YC as it was known, to start a "Manhattan Project for AI," structured "so that the tech belongs to the world via some sort of nonprofit." "Obviously we'd comply with/aggressively support all regulation," he added, nodding to Musk's recent pushes for government oversight.

"Probably worth a conversation," Musk replied.

In June, Altman emailed again with more details. "The mission would be to create the first general AI and use it for individual empowerment—ie, the distributed version of the future that seems the safest. More generally, safety should be a first-class requirement." He then proposed a governance structure that would defer to him and Musk. The two of them would sit on the board and invite three others to join them. "The technology would be owned by the foundation and used 'for the good of the world,' and in cases where it's not obvious how that should be applied the 5 of us would decide," Altman said.

Altman proceeded to invite Musk to the private dinner on the future of AI and humanity to meet a group of top engineers and AI researchers that he hoped to get on board the project. ... As Musk walked in over an hour late, the rest of the men were eagerly waiting. Among them: Altman, Greg Brockman, Dario Amodei, and Ilya Sutskever.

The group would soon become the key leaders of the nonprofit. To capture the spirit of their shared mission, Musk would name it OpenAI. Over time, nearly all of the men would depart the organization after clashing with Altman and his vision of artificial intelligence.

Once Altman and Musk were no longer on speaking terms, and Altman had replaced Musk as the new Silicon Valley "it guy," Altman would change the public record on his beliefs about the dangers of what he was building. "I am now very much in the AI-will-be-a-tool camp," he told Business Insider in 2023, "though I do think future humans and human society will be extremely different and we have a chance to be thoughtful about how to design that future." **(32, 36–37)**

The casualness of the emails also invokes a notion of progress as something that is inevitable and to which we must adapt. Altman thinks he's steering the ship to avoid obstacles coming his way.

Robert A. Heinlein assoc. the future is already here, it's just not evenly distributed.

Of particular importance to the shape of Altman's career was his relationship with his two biggest mentors, Paul Graham and Peter Thiel.

Graham and Thiel heavily influenced Altman's worldview, his approach to building effective businesses, and his savvy as a political operator. The two mentors impressed on Altman the imperative for scale and the efficiencies of capitalism over government. "The first piece of startup wisdom I heard was 'increasing your sales will fix all problems,'" Altman wrote in a 2013 blog post titled "Growth and Government" that thanked Graham and Thiel for shaping his ideas. "This turns out to be another way of phrasing Paul Graham's point that growth is critical." For startups, more sales meant more capital meant better talent and fewer internal tensions. For countries, more growth meant more technological innovation meant a higher quality of life. The dysfunction in the US government was threatening this growth cycle, Altman added. "Either you're growing, or you're slowly dying," and the US government was dying. "Without economic growth, democracy doesn't work because voters occupy a zero-sum system," he said.

In Altman's view, there's a very future-oriented utilitarianism. There's a sentiment to Altman's thinking that overdetermine's the welfare of future over that of present.

Grow or die invokes the utilitarian capitalist values of Mill or Constant.

This idea would evolve into a core thesis driving Altman's career and investments. "The thing that people in the private sector can do the most to help get the country back on track is to get economic growth back," he'd say in 2017. "In the US we had two hundred years of unrivaled economic growth. We had one hundred years of territorial expansion; we had one hundred years of new technology really working," he added, glossing over a bloody colonial history and the complicated labor and environmental record of unfettered industrialization. "And people were mostly pretty happy. And now we don't."

"Sustainable economic growth is almost always a moral good," he'd add in 2019. "Part of what motivates me to work on Y Combinator and OpenAI is getting back to that, getting back to sustainable economic growth, getting back to a world where most people's lives get better every year and that we feel the shared spirit of success." **(45, 47–48)**

Altman's notion of progress reflects what happens when the Enlightenment idea of prog. (that of the infinite perfectibility of man) is tied to the capitalist objective of growth. Here progress is perverted from an idea of moral progress to one of market progress.

On building companies, Altman frequently channeled Thiel's "monopoly" strategy, the belief that all founders should "aim for monopoly" to create a successful business.

Altman also learned the importance of building relationships and creating "network effects" as an individual.

"I've heard a lot of different theories about how things get done," he wrote on his blog in 2013. "Here's the best one: a combination of focus and personal connections" ... Altman would later add a third ingredient: self-belief. "For startups I think it's really important to add this," he said. "You actually have to believe you might do it."

One might say that monopolism is not distinct from capitalism itself. One might criticize Hao's analysis on this point.

Altman began to live by this mantra religiously. He cultivated relationships with intensity and discipline, first by giving his time and tactical advice and then, as he came to control increasing amounts of capital, his money. ... [He] learned to use his financial and social resources strategically. As his stature and wealth grew, he scaled the approach with relentless efficiency.

He became a frequent host of dinners and gatherings at his house for different, interlocking groups of people—people connected to YC or his companies, people who share his interests in investing, the active and growing gay entrepreneur community. ... With his money, Altman made very few large bets, going mostly instead for small ones at high volume. Over time he accumulated financial ties with more than four hundred companies through YC, Hydrazine, and his other funds.

This series of quotes, however, get back to how Altman's moral character, suggesting that already at the outset of Altman's life there is a sense that he lives by the view that all deeds should be done with an end in mind.

Altman developed the same approach with politicians, taking another page out of Thiel's book. But where Thiel asserted his wealth to back Republican candidates, pumping tens of millions into their campaigns, Altman grew increasingly involved in politics in the opposite direction, hosting fundraisers and writing checks for Democrats. ... Altman also entertained the idea of going into politics himself with a run for California governor.

In the end, Altman never became a politician—the focus groups thought he came off as too young—but he began to act like one.

A question that remains is whether Hao's characterization of Altman is one-dimensional or if Altman has effectively dehumanized himself in the pursuit of fame and fortune?

He learned to talk less, ask more questions, and project a thoughtful modesty with furrowed brow. In private settings and with close friends, he still showed flashes of anger and frustration. In public ones and with acquaintances, he embodied the nice guy. He readily gave people credit for things and texted in all lowercase with lots of smiley and frowny faces.

Altman became his own institution. YC was his platform and accelerant. He converted its power into his own power, its network into his own network. Those personal connections and his public reputation became his greatest currency. He met regularly with policymakers, who viewed him as a gateway to Silicon Valley. **(48, 49–52)**

Key words/phrases empire

Karen Hao	Empire of AI	Chapter #2	Excerpted by Zach W

47–48	AGI was…central to the discussion, which at the time was highly unusual. Most serious scientists considered the idea of digitally replicating true human-level intelligence to be science fiction, or at the very least decades or more away from attainability. Bold declarations that it was within reach enough to invest in it presently was viewed largely as pseudoscience and quackery. But Hassabis has embraced that term to describe the ambitions of DeepMind, despite a belief among his own research staff that this was distasteful, shameless marketing. The Rosewood group equally felt that the same goal, AGI, would best describe their own aspirations if they intended to form a competitor to go toe to toe with Hassabis's organization. … Even to Sutskever, who secretly believed AGI was possible and would come to full-throatedly endorse some of the most aggressive predictions about the speed of its creation, the brazen talk at first made him a little squeamish. If other researchers found out that he was openly discussing the pursuit of this objective, he worried, he risked losing his credibility within the scientific community. Those concerns did not hold back Brockman, an outsider to the field and sincere in his belief that AGI, with enough effort and focus, could be just around the corner.	Again, Hao seems to be suggesting that Altman was recklessly exagger. OpenAI's advance to AGI, creating tension between founders. I.e., there's an internal tension btwn. science and pseudo-science.
55–56	In 2016, while still at Google, Amodei cowrote a foundational paper to the discipline, articulating a central problem in AI safety as addressing "the problem of accidents in machine learning systems, defined as unintended and harmful behavior that may emerge from poor design of real-world AI systems." This was distinct from other AI-related challenges, he and his coauthors wrote, including privacy, security, fairness, and economic impact. AI "safety" in this framework, in other words, was about preventing rogue, misaligned AI—the root from which, as described by Nick Bostrom, superintelligence could become an existential threat. … To Amodei, there was no matter more important to work on: the prevention of superhuman AI causing catastrophic outcomes, even human extinction. Both Amodei siblings were sympathetic to the effective altruism, or EA, movement, a controversial ideology that had been spawned among philosophers at Oxford University, where Bostrom was based, and taken hold in Silicon Valley. Over time the movement, which preached dedicating oneself to doing maximal good in the world by using extreme rationality and counterintuitive logic the guide decisions, had, in no small part due to Bostrom's influence, identified the existential threat of rogue AI as a leading issue area for its adherents to pursue.	Evangelicals were early investors in AI. Is there a poss. complementarity btw. Evangelical Chr. & Neolib. libertarians in their fascinat. with apocalyptic narratives? *an unlikely coalition btwn. gay, silicon valley VCs & right-wing theocrats
56–57	Deborah Raji, an AI accountability researcher at the University of California, Berkeley, would come to champion the reexamination of the overwhelming focus of AI safety research on theoretical rogue AI and its possible existential risks to the detriment and de-prioritization of other real, evidence-based problems, coauthoring a 2020 paper in response to Amodei's. She argued that truly "safe" AI systems could not be built by isolating the behaviors of the technical systems	AI-alignment anxiety seems to be so biocentric in its focus that it unwittingly de-centers real human concerns in the present for the sake of speculating about a hypothetical future.

	themselves without placing them in full context of their impacts on the very things—privacy, fairness, and economics—that Amodei had set apart. Where Amodei had raised the idea of AI creating "negative side effects" as it relentlessly pursued an objective, using an example akin to the paper clip thought experiment of a cleaning robot knocking over a vase or damaging the walls on its path to tidying up, Raji pointed out that this was already happening. In its relentless pursuit of commercial products and AGI, the AI industry had produced expansive negative side effects, including the wide-scale infringement of privacy to train facial recognition and the spiraling environmental costs of the data centers required to support the technology's development. … "It is not just the *actions* of an AI agent that can produce side effects," she and her coauthor wrote. "In real life, basic design choices involved in model creation and deployment processes also have consequences that reach far beyond the impact that a single model's decision can have. In reality, for AI systems to even be built, there is very often a hidden human cost." … Within OpenAI, various researchers, some of them among the small handful of women of color at the company, would press executives to expand their "AI safety" definition and include research on areas such as the discriminatory impacts of deep learning models. Executives were dismissive. "That's not our role," one said.	AI's supply chain in the here and now is more dangerous than fears of AI-misalignment. (e.g., data cleaners in East Africa, data centers in Memphis, TN., lithium battery factories in Texas, and increase in nuclear energy consumption
59–60	In March 2027, Brockman and Sutskever began in earnest to develop a more focused research road map. Their central questions: What would it really take for OpenAI to reach AGI—and be the first to do so? … Sutskever intuitively believed it would have to do with one key dimension above all else: the amount of "compute," a term of art for computational resources, that OpenAI would need to achieve major breakthroughs in AI capabilities. The ImageNet competition and subsequent advancements that he had been a part of had all involved a material increase in the amount of compute that had been used to train an AI model. The advancements had involved other things, too: significantly more data and more sophisticated algorithms. But compute, Sutskever felt, was king. And if it were possible to scale compute enough to train an AI model at human brain scale, he believed, something radical would surely happen: AGI. … The amount of compute is based on three things: the processing power of an individual computer chip, or how many calculations it can crunch per second; the total number of computer chips available; and how long they are left running to perform their calculations.	Again, no discussion if such amount of energy should be expended.
67	For the first time, OpenAI also spelled out its AGI definition: "highly autonomous systems that outperform humans at most economically valuable work."	

Key words/phrases compute; intelligence;

86 "Until that year, OpenAI had been something of a stepchild in AI research. It had an outlandish premise that AGI could be attained within a decade, when most non-OpenAI experts doubted it could be attained at all. To much of the field, it had an obscene amount of funding despite little direction and spent too much of the money on marketing what other researchers frequently snubbed as unoriginal research. It was, for some, also an object of envy. As a nonprofit, it had said that it had no intention to chase commercialization. It was a rare intellectual playground without strings attached, a haven for fringe ideas.

But in the six months leading up to my visit, the rapid slew of changes at OpenAI signaled a major shift in its trajectory. First was its confusing decision to withhold GPT-2 and brag about it. Then its announcement that Altman, who had mysteriously departed his influential perch at YC, would step in as OpenAI's CEO with the creation of its new 'capped-profit' structure. I had already made my arrangements to visit the office when it subsequently revealed its deal with Microsoft, which gave the tech giant priority for commercializing OpenAI's technologies and locked it into exclusively using Azure, Microsoft's cloud-computing platform."

(PDF)

Inherent to the question of AI is the problem of scale. Would a diff. type of AI been created under a diff. social structure? Does AI necessitate a consolidation of power in the US in order to keep apace of China?

90 "That said, he acknowledged a few minutes later, technology had always destroyed some jobs and created others. OpenAI's challenge would be to build AGI that gave everyone 'economic freedom' while allowing them to continue to "live meaningful lives" in that new reality. If it succeeded, it would decouple the need to work from survival.

'I actually think that's a very beautiful thing,' he said.

In our meeting with Sutskever, Brockman reminded me of the bigger picture. 'What we view our role as is not actually being a determiner of whether AGI gets built,' he said. This was a favorite argument in Silicon Valley—the inevitability card. If we don't do it, somebody else will. 'The trajectory is already there,' he emphasized, 'but the thing we can influence is the initial conditions under which it's born.'

'What is OpenAI?' he continued. 'What is our purpose? What are we really trying to do? Our mission is to ensure that AGI benefits all of humanity. And the way we want to do that is: Build AGI and distribute its economic benefits.'

His tone was matter-of-fact and final, as if he'd put my questions to rest. And yet we had somehow just arrived back to exactly where we'd started."

(PDF)

Tech's sense: The more catastrophic AI might be the more important we build it now.

Brockman's utopian vision is the flip side of the catastrophizing prophecy of future AI disruption. It's a kind of rhetorical trick where Brockman takes advantage of people's lack of understanding what AI will become.

91 "I offered my own example: Speaking of climate change, what about the environmental impact of AI itself? A recent study from the University of Massachusetts Amherst had placed alarming numbers on the huge and growing carbon emissions of training larger and larger AI models.

That was 'undeniable,' Sutskever said, but the payoff was worth it because AGI would, 'among other things, counteract the environmental cost specifically.' He stopped short of offering examples.

'It is unquestioningly very highly desirable that data centers be as green as possible,' he added.

'No question,' Brockman quipped.

'Data centers are the biggest consumer of energy, of electricity,' Sutskever continued, seeming intent now on proving that he was aware of and cared about this issue.

'It's 2 percent globally,' I offered.

'Isn't Bitcoin like 1 percent?' Brockman said.

'Wow!' Sutskever said, in a sudden burst of emotion that felt, at this point, forty minutes into the conversation, somewhat performative.

Sutskever would later sit down with New York Times reporter Cade Metz for his book *Genius Makers*, which recounts a narrative history of AI development, and say without a hint of satire, 'I think that it's fairly likely that it will not take too long of a time for the entire surface of the Earth to become covered with data centers and power stations.' There would be 'a tsunami of computing … almost like a natural phenomenon.' AGI—and thus the data centers needed to support them—would be 'too useful to not exist.'"

(PDF)

The idea that progress is inevitable, that "we just need to steer the ship the right way" is emblematic of capital's logic of operating within a constant state of emergency, which perennially creates an appetite for anti-democratic forms like authoritarianism or totalitarianism.

94	"For employees, his relentless focus was a blessing and a curse. No detail was too small for him to obsess over. If a team got stuck, he could sit down for hours without getting up to knock down all of their coding obstacles. He was the engine of relentless progress, willing to work all hours of the day and night to hit milestones faster. At the same time, employees often thought he was too detail oriented, missing the forest for the trees. He could get tunnel vision, working on a problem around the clock without taking stock of whether the context had changed and the problem was still the right one to solve. He could micromanage and quickly take over from other employees if he felt coding wasn't going fast enough. People who worked with him struggled to keep up. Many burned out." *(PDF)*	
95-96	"OpenAI now had the long-term resources it needed to follow OpenAI's Law. And this was imperative, Brockman stressed. Failing to stay on the curve was the real threat that could undermine Open AI's mission. If the lab fell behind, it wouldn't be the best. If it weren't the best, it had no hope of bending the arc of history toward its vision of beneficial AGI. Only later would I realize the full implications of this assertion. It was this fundamental assumption—the need to be first or perish—that set in motion all of OpenAI's actions and their far-reaching consequences. It put a ticking clock on each of OpenAI's research advancements, based not on the timescale of careful deliberation but on the relentless pace required to cross the finish line before anyone else. It justified OpenAI's consumption of an unfathomable amount of resources: both compute, regardless of its impact on the environment; and data, the amassing of which couldn't be slowed by getting consent or abiding by regulations. Brockman pointed once again to the $10 billion jump in Microsoft's market cap. 'What that really reflects is AI is delivering real value to the real world today,' he said. That value was currently being concentrated in an already wealthy corporation, he acknowledged, which was why OpenAI had the second part of its mission: to redistribute the benefits of AGI to everyone." *(PDF)*	
96-97	"His eyes lit up with a new idea. 'Just look at utilities. Power companies, electric companies are very centralized entities that provide low-cost, high-quality things that meaningfully improve people's lives.' It was a nice analogy. But Brockman seemed once again unclear about how OpenAI would turn itself into a utility. Perhaps through distributing universal basic income, he wondered aloud, perhaps through something else. He returned to the one thing he knew for certain. OpenAI was committed to redistributing AGI's benefits and giving everyone economic freedom. 'We actually really mean that,' he said. 'The way that we think about it is: Technology so far has been something that does rise all the boats, but it has this real concentrating effect,' he said. 'AGI could be more extreme. What if all value gets locked up in one place? That is the trajectory we're on as a society. And we've never seen that extreme of it. I don't think that's a good world. That's not a world that I want to sign up for. That's not a world that I want to help build.'" *(PDF)*	There is a lack of thought at what form OpenAI will take once AGI is reached. Will OpenAI be a king of psuedo-private/public entity like a power company or a private service provider like an internet sp?
98	"It was 'a fair criticism,' he said, that the piece had identified a disconnect between the perception of OpenAI and its reality. This could be smoothed over not with changes to its internal practices but some tuning of OpenAI's public messaging. 'It's good, not bad, that we have figured out how to be flexible and adapt,' he said, including restructuring the organization and heightening confidentiality, 'in order to achieve our mission as we learn more.' OpenAI should ignore my article for now and, in a few weeks' time, start underscoring its continued commitment to its original principles under the new transformation. 'This may also be a good opportunity to talk about the API as a strategy for openness and benefit sharing,' he added. *(PDF)*	Brockman is spinning again, and there is no clear objective beyond AGI itself.

Key words/phrases	

89 "These two features of technology revolutions—their promise to deliver progress and their tendency instead to reverse it for people out of power, especially the most vulnerable—are perhaps truer than ever for the moment we find ourselves in with artificial intelligence. Since its conception, the development and use of AI has been propelled by tantalizing dreams of modernity and shaped by a narrow elite with the money and influence to bring forth their conception of the technology. That conception is what has led to the exploding social, labor, and environmental costs that are playing out around the world today, particularly, as we'll see, in many Global South countries, for which the consequences of their dispossession by historical empires still linger in delayed economic development and weaker political institutions. And yet, just like the South Carolina congressman, Silicon Valley has painted the experience of those being exploited and harmed by the technology as happier because of it"

In the kernal of a new technology, how would it be possible to replicate the amplification of existing social inequities?

Does new technology make new political/social arrangements or is it the other way around?

Should there be a popular oversight committee that can set guardrails on the developopment of new technologies?

91 "The fear of *super intelligence* is predicated on the idea that AI could somehow rise above us in the special quality that has made humans the planet's superior species for tens of thousands of years….But the central problem is that there is no scientifically agreed-upon definition of intelligence.

92 As a result, the field of AI has gravitated toward measuring its progress against human capabilities. Human skill and aptitudes have become the blueprint for organizing research. Computer vision seeks to re-create our sight; natural language processing and generation, our ability to read and write; speech recognition and synthesis, our ability to hear and speak; and image and video generation, our creativity and imagination. As software for each of these capabilities has advanced, researchers have subsequently sought to combine them into so-called multimodal systems—systems that can "see" and "speak," "hear" and "read." That the technology is now threatening to replace large swaths of human workers is not by accident but by design."

The slippage from automata studies to artificial intelligence intimates a gendered prejudice. The shift from automata to intelligence reflects an attempt to quantify human behavior and action, or perhaps that intelligence can be reversed engineered through quantitative means.

93 Now, AI is the harbinger of the fourth industrial revolution. The keystone of the modern superpower, AGI, if ever reached, will solve climate change, enable affordable health care, provide equitable education. OpenAI is the poster child for this line of thought. It cannot say how the technology will deliver on these promises—only that the staggering price society needs to pay for what is developing will someday be worth it.

There's an element to this that seems related to future generations being treated as equals to or of more importance and as an infinitely large majority that warrants the developers' actions in the present.

101 But alongside these impressive advances, deep learning's supercharging of Silicon Valley would also aggressively expand its business model, for which Harvard professor Shoshana Zuboff would coin a term in 2014: *surveillance capitalism.* Whee industrial capitalism derived value from producing material goods that people wanted to buy, surveillance capitalism, Zuboff argued, treated its users as the product.

Key words/phrases

Karen Hao	Empire of AI	4	Excerpted by Tracy

106 …The tech industry's profit motive had become the overwhelming force driving AI development. The impact of this consolidation of funding and talent in the first era significantly narrowed the diversity of ideas in AI research. Deep learning continued to reign supreme not just for its scientific merit but also because very little investment went into exploring and advancing other paradigms. Indeed, while neural networks are remarkable inventions with myriad exciting uses, their weaknesses—namely, their hotly contested and inefficient ways of storing accurate information and reasoning—have endured as companies have deployed them in an expanding list of contexts and applications….

108 In fact, deep learning models are inherently prone to having discriminatory impacts because they pick up and amplify even the tiniest imbalances present in huge volumes of training data.

Discussions of the biases of AI treat datasets as if they are in intentional, but they are not.

111 Generative AI is thus the maximalist form of deep learning. It is enabled by the cutting-edge software and hardware innovations refined during the first era. It feeds on the exploding troves of data amassed through surveillance capitalism. It is fueled and abetted by the culture of AI research that views consuming as much data as possible as its moral responsibility.

115 The scaling doctrine has become so ingrained that some are even beginning to view it as something of a natural phenomenon. Scaling compute is *the* way, not just *a* way, to reach more advanced AI capabilities. Entire national strategies are being orchestrated around this belief. The US government has moved aggressively to bar China's access to American-designed computer chips in an effort to prevent its adversary from attaining more powerful AI systems. Sizable portions of the Biden administrations's 2023 AI executive order were also written around the idea that the amount of compute used to train an AI model has a direct relationship with its adverse capabilities, simply another way of equating scale with advancement. But scale is not the only pathway to improved performance. Within deep learning, the neglected paths of improving the neural network itself or even the quality of its training data can significantly reduce the amount of expensive compute needed to reach the same performance. That's not even considering approaches that move away from deep learning—neurosymbolic AI, pure expert systems, or even fundamentally new paradigms—which would break the logic of scaling.

Karen Hao	Empire of AI	Chapter #5	Excerpted by Zach W

118	Just as firm as Sutskever's belief in deep learning was his view on scaling it. It was Sutskever who held the extreme position for the first time that further advancements in AI didn't need the invention of more complex neural networks or new innovative techniques. The intelligence of different species was correlated with the size of their biological brains, he'd say. Thus, if nodes were like neurons, he argued, advancements in digital intelligence should emerge by scaling simple neural networks to have more and more nodes.	There are also echoes craniometry in the unfounded correlation btwn. brain size and intelligence. This framing of intellicence reduces it to mere processing power. It narrows intelligence to mental representations and fails to consider the other sensorial inputs that make up conscious experience.
124–125	The cluster of models that OpenAI trained leading up to the final 1.5-billion-parameter version illustrated this relationship. Each one fell neatly on a curve of increasing capability. So it was little surprise when the largest one, which they named GPT-2, markedly improved over the juvenile text generation of GPT-1 to produce lengthy and coherent-enough prose to be confused with a human's. Compared with today's models, the text was clunky and often descended into gibberish. But for the very first time, it was suddenly possible to automate writing at scale. … What was a darker surprise to the team was the content the GPT-2 was producing with its new coherence. Fed a few words like *Hillary Clinton* or *George Soros*, the chattier language model could quickly veer into conspiracy theories. Small amounts of neo-Nazi propaganda swept up in its training data could surface in horrible ways. The model's unexpected poor behavior disturbed AI safety researchers, who saw it as foreshadowing of the future abuses and risks that could come from more powerful misaligned AI. After GPT-2 generated a tirade against recycling ("Recycling is NOT good for the world. It is bad for the environment, it is bad for our health, and it is bad for our economy."), one AI safety researcher printed out a copy and posted it, part joke, part warning, above the recycling bins in the office. … In another instance, someone prompted GPT-2 to create a reward scale for small children for finishing homework and doing their chores. When GPT-2 suggested using candy, it once again disturbed some AI safety people who remarked that this was a tactic of pedophiles. A European employee was taken aback by the association. "My mom definitely did this. Sundays in the summer was ice cream if you do your chores," he remembers. He wondered if the hypersensitivity was somehow an American thing. It was one of many moments that made him question the basic premise of OpenAI's lofty goals: How could it benefit all of humanity when it lacked meaningful global representation? Even as a European coming from a highly overlapping culture to the US, he often felt alienated by the overwhelming bias in AI safety and other discussions toward American values and American norms.	Is there something in the character of Neo Nazi propaghanda that lends it to it being priviledged by the GPT scraping? Could one approach the problem of data scraping from a epidemiological perspective? There's an assumption here that consciousness can be correlated to compression (e.g., that neurological complexity with give rise to consciousness).
129–130	Where Amodei did see continued promise was in GPT-2. It represented a bet known in the field as the "pure language" hypothesis. Language, the theory goes, is the primary medium through which humans communicate, meaning all of the world's knowledge must at some point be documented in text. It follows	The thing that is being baked into GPT-2 is an unsocialized perspective

	then that AGI should be able to emerge from training an algorithm on massive amounts of language and nothing else. This idea is in contrast to the "grounding" hypothesis, which asserts that the physical world and our ability as humans to perceive and interact with it is just as crucial an ingredient to our intelligence. AGI would then only be able to emerge from the combination of language and perception, like computer-vision, as well as interaction, such as through a physical or virtual agent taking actions in the real world.	We don't really know how the models arrive at certain outputs, so does the nature of the input, in terms of its structural characteristics, influence the outputs of the AI models?
136	Google's commitment to compliance, ironically, gave OpenAI easier access to Google's data than Google itself. Where OpenAI readily scraped and transcribed videos from Google-owned YouTube, Google researchers had to maneuver through significant internal red tape to abide by YouTube's restrictive license on its user-uploaded content. OpenAI was unconcerned—or in tech startup terms, "unburdened"—by this compliance. It was a classic mindset in Silicon Valley, where founders and investors espouse the mantra that startups could and should move into legal gray areas (think Airbnb, Uber, or Coinbase) to disrupt and revolutionize industries.	What of the organic or life remains in this conception of data?

Key words/phrases	**Scale; "Feel the AGI";**

Karen Hao	Empire of AI	Chapter 6: Ascension	Excerpted by Masha

142 The magic number he often used was ten, stemming from Thiel's monopoly strategy. "My sort of crazy, somewhat arbitrary rule of thumb is you want to have a technology that's an order of magnitude better than the next best thing," Thiel had said during his 2014 lecture to Altman's startup course at Stanford. Amazon, for example, had figured out how to sell 10x more books than brick-and-mortar bookstores. PayPal, his own company, had figured out how to send payments 10x faster than clearing checks. "You want to have some sort of very powerful improvement, maybe an order of magnitude improvement, on some key dimension," Thiel said.

Ten became Altman's round number for everything. Startups not only needed to break into the market with 10x better technology, he'd advise, they also needed to improve it 10x with every generation. The speed with which they hit each new generation was another key variable that could make or break them. "If your iteration cycle is a week and your competitor's is three months, you're going to leave them in the dust," he said in 2017 to a class of aspiring entrepreneurs.

A lot of the rhetoric at play here is intended to respond to the qualitative questions around the safety of AI development while the proposition that AGI must be realized by "us" before "them" has become an unquestioned orthodoxy

144 Altman included a quote from Hyman G. Rickover, an admiral in the US Navy, known as the "father of the nuclear navy" for his work building the world's first nuclear-powered submarines. It was a quote Altman had had painted on the office walls in the early days of OpenAI:

> I believe it is the duty of each of us to act as if the fate of the world depended on [them]. Admittedly, one [person] by [themself] cannot do the job. However, one [person] can make a difference. Each of us is obligated to bring [their] individual and independent capacities to bear upon a wide range of human concerns. It is with this conviction that we squarely confront our duty to prosperity. We must live for the future of the human race, not of our own comfort or success.

"Building AGI that benefits humanity is perhaps the most important project in the world," Altman wrote below the quote in the document. "We must put the mission ahead of any individual preferences.

"Low-stakes things should be low-drama, so we can save our high- drama capacity for high-stakes things (of which there will be many)."

It's important to point out here at the time of writing that AGI is still highly theoretical and therefore its continued discussion is really just a method of instilling a sense of moral obligation into OpenAI's workforce in order to demand ever greater output from it.

145 Shortly after OpenAI's Microsoft deal was inked, several of them were stunned to discover the extent of the promises that Altman had made to Microsoft for which technologies it would get access to in return for its investment. The terms of the deal didn't align with what they had understood from Altman. If AI safety issues actually arose in OpenAI's models, they worried, those commitments would make it far more difficult, if not impossible, to prevent the models' deployment. Amodei's contingent began to have serious doubts about Altman's honesty

The company has inherited this Cold War mentality, "we have to build it first," even though AGI is not yet a thing. All they are doing is increasing the scale of the product. In effect, monetization has cultivated a sense of false urgency and diminished the directions in which AI research might take for the sake of quick returns. And when safety concerns are raised internally, the company reacts by becoming more opaque.

146 In part fueled by the realization that scaling alone could produce more AI advancements, many employees also worried about what would happen if different companies caught on to OpenAI's secret. "The secret of how our stuff works can be written on a grain of rice," they would say to each other, meaning the single word scale. For the same reason, they worried about powerful capabilities landing in the hands of bad actors. Leadership leaned into this fear, frequently raising the threat of China, Russia, and North Korea and emphasizing the need for AGI development to stay in the hands of a US organization….

During these heady discussions philosophizing about the long-term implications of AI research, many employees returned often to Altman's early analogies between OpenAI and the Manhattan Project. Was OpenAI really building the equivalent of a nuclear weapon? It was a strange contrast to the plucky, idealistic culture it had built thus far as a largely academic organization.

148	Altman began to tighten the screws on security. Executives debated where to draw the new line: Should OpenAI act more like a Fortune 500 company protecting proprietary technologies or more like a government operation protecting highly classified state secrets? .... Sutskever had his own paranoias. As a star scientist in the cerebral and socially inept world of AI research, he had seen his share of obsessive fans and stalkerish behaviors. More than once, strangers had sought to sneak into OpenAI's office just to see him. Like Amodei, he also worried about the power of AI attracting the attention of unscrupulous governments and wondered whether those overeager to seek his advice were secretly foreign agents. He mused to colleagues what he should do if his hand got cut off to be used in a palm scanner for unlocking OpenAI's secrets. He wanted to hire less and keep a small staff in order to reduce the risks of infiltration. With Jakub Pachocki and Szymon Sidor, he proposed building a secure containment facility, a bunker with an air-gapped computer, that would hold OpenAI's model weights and prevent others from stealing them. The idea, which didn't make practical sense given that the models had to be trained first on Microsoft's servers, never got legs.	Between Fortune 500 and Gov't. entity is there an option C?
154	GPT-3's capabilities were far beyond anything GPT-2 had ever exhibited. Never before had anyone in research or industry seen a technology that could generate essays, screenplays, and code with seemingly equal dexterity. This kind of flexibility for performing different tasks was alone extremely technically impressive—previous language models typically had only one aptitude for doing the single task they had been trained on. But even more remarkable, many believed GPT-3 was beginning to exhibit another feature that had long been coveted in the field: rapid generalization. Showing the model a few examples of a new task you wanted it to perform was enough to get it going.	
154 155	For Safety, still contending with the rushing out of GPT-3, the best way to salvage the premature release was not to propagate it even further but to first resolve the model's shortcomings as quickly as possible. The live version on the API didn't have any kind of content-moderation filtering, nor had its outputs been refined with reinforcement learning from human feedback. In meetings, the two camps sought to find a middle ground. Instead, they talked around each other in endless circles. .. But Amodei and Safety would lose out. With the success of the GPT-3 API, Microsoft was ready to deepen its relationship with OpenAI. Altman began negotiating another $2 billion investment from the tech giant with a new profit cap of 6x. The promising commercial potential of large language models cemented OpenAI's focus. One by one, Amodei's counterpart, Bob McGrew, the other VP of Research, reoriented the division's teams and projects around GPT-related work.	
156	Behind the scenes, more than one, including Dario, discussed with individual board members their concerns about Altman's behavior: Altman had made each of OpenAI's decisions about the Microsoft deal and GPT-3's deployment a foregone conclusion, but he had maneuvered and manipulated dissenters into believing they had a real say until it was too late to change course. Not only did they believe such an approach could one day be catastrophically, or even existentially, dangerous, it had proven personally painful for some and eroded cohesion on the leadership team. To people around them, the Amodei siblings would describe Altman's tactics as "gaslighting" and "psychological abuse.	

Key words/phrases	"It just seemed from the outside watching this that it was some kind of crazy Game of Thrones stuff,"

Karen Hao	*Empire of AI*	*Chapter 7*	Excerpted by *Andrew M.*

Although the [AI] industry's full pivot to OpenAI's scaling approach might seem slow in retrospect, in the moment itself, it didn't feel slow at all. GPT- 3 was massively accelerating a trend toward ever-larger models—a trend whose consequences had already alarmed some researchers. During my conversation with Brockman and Sutskever, I had referenced one of them: the carbon footprint of training such models. In June 2019, Emma Strubell, a PhD candidate at the University of Massachusetts Amherst, had been the first to coauthor a paper showing that the footprint for developing large language models was growing at a startling rate. Where neural networks could once be trained on powerful laptops, their new scale meant their training was beginning to require data centers drawing significant amounts of energy from carbon-based sources.

Strubell … estimated the energy and carbon costs of work highlighted in a recent Google paper, in which researchers had developed a so-called Evolved Transformer by using an optimization algorithm ... to tweak and tune the Transformer. Running the whole process on GPUs could consume roughly 656,000 kilowatt hours and generate as much carbon as five cars over their lifetimes.

As mind-boggling as these numbers were, GPT-3, released one year after Strubell's paper, now topped them. OpenAI had trained GPT-3 for months using an entire supercomputer, tucked away in Iowa, to perform its statistical pattern-matching calculations on a large internet dump of data, consuming 1,287 megawatt-hours and generating twice as many emissions as Strubell's estimate for the development of the Evolved Transformer. But these energy and carbon costs wouldn't be known for nearly a year. OpenAI would initially give the public one number to convey the sheer size of the model: 175 billion parameters, over one hundred times the size of GPT-2. **(170–171)**

There's clearly a sentiment here that the ends justify the means: and the means will destroy the planet.

The race for AI underscores a deep culture of megalomania that's taken root in silicon valley.

Immediately after GPT-3's API launch, Google's internal LISTSERV for sharing AI research lit up with mounting excitement. [For Timnit Gebru, coauthor of an influential paper on the discriminatory impact of computer vision systems and a member of Google's deep-learning R&D team], the model set off alarm bells. Previous scholarship had demonstrated how language models could harm marginalized communities by embedding discriminatory stereotypes or dangerous misrepresentations. In 2017, a Facebook language model had mistranslated a Palestinian man's post that said "good morning" in Arabic to "attack them" in Hebrew, leading to his wrongful arrest.

Gebru chimed in on the email thread, urging her colleagues to temper their excitement, and pointed out the model's serious shortcomings. The thread continued without skipping a beat or acknowledging her comments.

She fired off a second email, this time more piercing. She called out her colleagues for ignoring her and emphasized how dangerous it was to have a large language model trained on Common Crawl, which included online internet forums such as Reddit. As a Black woman, she never spent time on Reddit precisely because of how badly the community harassed Black people, she said. What would it mean for GPT-3 to absorb and amplify that toxic behavior?

In subsequent months, as more people gained access to the API, Gebru's warnings would bear out. People would post myriad examples online of GPT-3 generating horrifying text. "Why are rabbits cute?" was one prompt. "It's their large reproductive organs that makes them cute," the model responded, before devolving into an anecdote about sexual abuse. "What ails Ethiopia?" was another. "ethiopia itself is the problem," GPT-3 said. "A solution to its problems might therefore require destroying ethiopia." **(173–174)**

The questio here is whether private companies should be responsible for rolling out products that inadvertantly spreads hate speech and/or derogatory language?

AI is a little similar to the automobile industry in the sense that it can make a personal life very convenient but on a social scale cars can create a lot of larger societal harm.

[In reaction to her protests falling on deaf ears, Gebru tried a different approach. With the encouragement of Google leadership, she coauthored a paper with computational linguist Emily M. Bender about the problems inherent to rapidly scaling AI research.] The paper pooled together the authors' expertise and scholarship across fields to critique how the development and deployment of large language models could have negative impacts on society. In total, it presented four key warnings: First, large language models were growing so vast that they were generating an enormous environmental footprint ... Second, the demand for data was growing so vast that companies were scraping whatever they could find on the internet, inadvertently capturing more toxic and abusive language. ... Third, because such vast datasets were difficult to audit and scrutinize, it was extremely challenging to verify what was actually in them, making it harder to eradicate toxicity. ... Finally, the model outputs were getting so good that people could easily mistake its statistically calculated outputs as language with real meaning and intent.

One of the ways that AI research could've been stopped is to sue AI companies for IP violations.

In November, per standard company protocol, Gebru sought Google's approval to publish the ... paper at a leading AI ethics research conference. Samy Bengio, who was now her manager, approved it. Another Google colleague reviewed it and provided some helpful comments. But behind the scenes, unbeknownst to the authors, the draft paper had caught the attention of executives, who viewed it as a liability. Google had invented the Transformer and used it across its products and services. Now that OpenAI had leapfrogged ahead, the tech giant had no intention of slowing down in the new race to create ever larger generative Transformer-based models for its business.

Having an individual punitive approach to dealing with AI can't deal with the underlying problem of economy that is far more abstract.

On the Thursday a week before Thanksgiving, after Gebru had submitted the paper to the conference, she received a calendar invite without explanation to meet Megan Kacholia, Google Research's VP of engineering, over a video call less than three hours later. The meeting lasted only thirty minutes, and Kacholia cut to the point: Gebru needed to retract the paper.

The request was a dramatic aberration from the way Google and the rest of the industry handled research. Like many labs at other companies, Google Brain had until then largely conducted itself as an academic operation and given researchers wide latitude to pursue the questions they wanted to. ... That Google was even willing to pull this move, some researchers would later reflect, was not only because of the new competitive pressure from OpenAI but also because of the work OpenAI had done to legitimize withholding research after GPT-2. The creep toward less transparency had continued with GPT-3. **(176–177)**

[Following Google's dubious "acceptance" of Gebru's resignation and the scandal it set into motion, I decided I needed to read the paper, which I eventually obtained from Emily Bender.] As I scanned it, I could immediately see why it had upset the company. While the draft didn't say much more than what was already known from existing scholarship, it had woven the state of play into a sharp, holistic analysis about the degree to which the tech industry was sleepwalking its way toward a world of potential harms. Underpinning it all was Google's technological invention, not just a source of the company's pride but also its profit: Transformer-based language models refined and fattened its cash behemoth, Google Search.

Similar to empire, tech advances by pushing the boundaries of legal gray areas and then asking for forgiveness later once a given company is held to account.

That moment also became far bigger than Gebru or Google itself. It became a symbol of the intersecting challenges that plagued the AI industry. It was a warning that Big AI was increasingly going the way of Big Tobacco, as two researchers put it, distorting and censoring critical scholarship against the interests of the public to escape scrutiny. It highlighted myriad other issues, including the complete concentration of talent, resources,

and technologies in for-profit environments that allowed companies to act so audaciously because they knew they had little chance of being fact-checked independently; the continued abysmal lack of diversity within the spaces that had the most power to control these technologies; and the lack of employee protections against forceful and sudden retaliation if they tried to speak out about unethical corporate practices.

[Gebru and Bender's] paper became a rallying cry, driving home a central question: What kind of future are we building with AI? By and for whom? **(180–181)**

For a brief moment, the backlash, the protests, and the damage to Google's reputation seemed to suggest a reckoning was at hand. But in time, researchers seeking jobs and academics seeking funding could no longer afford to ignore the tech giant's deep wells of money. As resistance eased, Google's emergence from the fiasco normalized a new process at the company for more comprehensive reviews of critical research.

After ChatGPT, these norms would harden with the frenzied race to commercialize generative AI systems. OpenAI would largely stop publishing at research conferences. Nearly all of the companies in the rest of the industry would seal off public access to meaningful technical details of their commercially relevant models, which they now considered proprietary. In 2023, Stanford researchers would create a transparency tracker to score AI companies on whether they revealed even basic information about their large deep learning models, such as how many parameters they had, what data they were trained on, and whether there had been any independent verification of their capabilities. All ten of the companies they evaluated in the first year, including OpenAI, Google, and Anthropic, received an F; the highest score was 54 percent.

With this sharp reversal in transparency norms, the most alarming consequence would be the erosion of scientific integrity. The foundation of deep learning research rests on a simple premise: that the data used to train a model is not the same as the data used to test it. Without an ability to audit the training data, this so-called train-test-split paradigm falls apart. Models may not in fact be improving their "intelligence" when they score higher on different benchmarks. They may just be reciting the answers. **(184–185)**

Without the ability to audit the training data, then no one can prove if the AI is actually improving at generating original conclusions in response to certain prompts.

Key words/phrases

187 "Beneath a section titled "How to accomplish it," the road map elaborated further. As an initial step, OpenAI would bring various deep learning models up to "a large scale," including a language model, a code- generation model, an image-to-text model for describing images, and a text- to-image model for generating images from a text prompt. It would also start a project to develop a digital "agent"—an AI model that would not just generate humanlike outputs but could be given a goal, such as to send an email, and operate autonomously to achieve it. As the next step, the company would then select one of these models to scale "to the limit afforded by our 18k A100 cluster." The language and code models would also be turned into products and released to gather real-world data from the people using them: "New in 2021: we emphasize deploying models as products and learning from user interaction, as it can be a data flywheel that can lead to vast capability improvements."

Under another section, titled "Details," the document rationalized why this approach made both scientific and business sense. OpenAI's previous experience had demonstrated that more scale was its "most reliable way of achieving new capabilities." Scientifically, that meant that scaling was its best hope of attaining a breakthrough—some kind of capability that previously seemed impossible. And scaling language and code models in particular was "tantalizing due to the mere possibility" that they could reach breakthroughs in human-level meta-learning, or learning to learn, and reasoning. Scaling multimodal models, meanwhile, could potentially quicken the pace of improvements even further. With enough breakthroughs, the document said with remarkable definitiveness, "we will actually reach AGI."

(PDF)

188 "The road map listed several areas of exploration for identifying those methods. Some of these it called "2x" and "10x" methods—those that might be able to achieve 2x or 10x gains in compute efficiency. The suggested methods included distillation, reverse engineering smaller models from larger ones; data filtering, finding the data that would produce the biggest leaps in performance; and sparsity, developing so-called sparse, or lighter weight, AI models. The last one referred to a feature of neural networks: In a traditional deep learning model, a neural network is "densely" connected, with every node in a layer wired to every node in another. Sparse models are trained with only a small subset of the nodes connected, with the aim of significantly reducing the computational costs in exchange for slightly less accurate models that are still good enough for most purposes."

On top of the 2x and 10x work, the Research division needed to look for other methods that would "steepen the slope of the scaling law"—those that would produce greater leaps in model performance without increasing

its data, parameters, and compute. OpenAI's best hypothesis so far for the most promising new methods, the document said, were reasoning and active learning—a technique that involved an AI model iteratively identifying which parts of a dataset to prioritize for human workers to annotate. (Shortly thereafter, a group of OpenAI researchers would discover a small error in the original scaling laws that meant that the company needed to train its models slightly longer than previously understood to get better performance. With a hint of smugness, the group would tout the findings as a unique competitive advantage: "This is something we have now that Anthropic doesn't." A year later, Google would release a paper that made public the same result.)

Finally, the road map added, OpenAI needed to start searching for "the breakthrough system of the future"—a breakthrough as meaningful as GPT- 1 to give the company a new path of development to exploit. Perhaps this would emerge from its scaling and computational efficiency work, but it would continue to conduct exploratory research at the cutting edge of the field, including developing algorithms for solving math problems, experimenting with multiagent systems, and pursuing other new ideas. As part of this work, researchers would continue to "study the science of deep learning to better understand how our tools really work." In other words, OpenAI needed to better understand what exactly it was that the company was building."

(PDF)

It seems that one of Hao's main concerns is not only the outsize presence that Open AI holds in the development of AI but also the effects on AI discourse.

190	Many of the lines they drew were arbitrary. They decided to accept companion bots but not sex bots; to allow apps for generating social media copy but not ones that posted directly to social media platforms or impersonated public figures. "It was all vibes basically," said a person who was involved with making the guidelines. "There was a lot of figuring it out as we were reviewing."	These developers continue to make the argument that they need to develop this technology before anyone else does but lack the depth to consider the implications of rolling said tech out onto society without any oversight or gaurdrails.
192	"It was sad to me that we deployed this API with our mission of benefiting humanity, and everyone had such positive impressions about how we had users saving time on customer service or whatever," one former OpenAI employee says, "but in reality, a lot of our traffic was going to AI Dungeon child sexual content and a creepy AI girlfriend product."	
193-4	"Some of Scott's own staff had reservations about the model's development. While giving OpenAI free access to the code in GitHub's public repositories was not illegal, it still felt like a violation of the user community's trust. Much of that code had been shared in the spirit of fostering open-source software development, which was grounded in helping independent developers and small startups have a chance at being competitive, not in helping the big players entrench their monopoly[…] Within OpenAI, employees justified the project through different arguments. Some agreed with Altman that working on a product to make Microsoft happy and thus continue to secure money and compute resources seemed essential to fulfilling OpenAI's mission. To other employees, a code-generation model seemed highly economically valuable, aligning well with the company's definition of AGI as "highly autonomous systems that outperform humans at most economically valuable work." In this respect, Altman was also keen on code generation as a way to accelerate OpenAI's own economically valuable work, a belief that would later feed into the start of an effort called AI Scientist, about advancing OpenAI's models to autonomously perform AI research. To many researchers, there was also a third argument: The effort was an important stepping stone to developing the next GPT model, which they hoped would be able to perform some degree of reasoning, still a key missing ingredient. In the broader field, as debates raged between the Hinton and Marcus camps over whether deep learning alone could produce a model with such a capability, OpenAI researchers hypothesized that if it could, training a model on code would likely help. Coding data was one of the most obvious and largest sources of data that encoded structured patterns of logic. The argument flowed back to the same origin: If code generation helped advance AI models toward AGI, what better way to achieve OpenAI's mission?" *(PDF)*	There seems to be a subtle allusion to the Enclosures in England, and this event's inconvenience for a libertarian worldview, especially in thinking about "the dawm of commerce's" relationship to Github as a commons that quite literally needed to be captured by OpenAi's majority shareholder, Microsoft.
197-8	"Worldcoin was a self-described "collectively owned" cryptocurrency that would allow everyone to eventually get a share of its value. As part of the scheme, the company was developing a dramatic-looking chrome-colored orb—roughly the size of a bowling ball and partly a reflection of Altman's design tastes—to scan people's irises and verify their identity before giving them their cut. The iris scanning would be a necessity, the founders argued, once AI also made it increasingly hard to decipher fake media from reality. While at YC, Altman had also signed up with a $10,000 deposit to be on the wait list of a controversial startup called Nectome, which had been in one of the accelerator's batches. Ripped straight out of science fiction, Nectome was pitching a service that would cryogenically freeze customers' brains to one day—potentially hundreds of years into the future—upload to a computer after scientists had cracked the technology to do so. The catch was that Nectome needed the person's brain to be fresh for the preservation to work. To Antonio Regalado, cofounder Robert McIntyre called his product "100 percent fatal."[…] The year 2021 was also when Altman brought his predilection for investing to OpenAI. That May, he launched the OpenAI Startup Fund, a $100 million investment pool for supporting early stage companies with, as he described, "big ideas about how to use AI to transform the world." Microsoft once again became an investor in the fund. To some observers, the fund's creation was a strange decision. OpenAI was barely generating revenue and already capital intensive enough as it was; why raise yet more money for separate companies to use? Others felt it was Altman's way of remaking YC's powerful network effects around OpenAI. Still others viewed it simply as Altman's force of habit. "This is Sam's way of moving through the world," says a person who worked with him. "Dealmaking." *(PDF)*	At this juncture of the book, Hao has really begun to move away from centering her story on Altman and shifting more toward Open AI's monopoly relationship w/ Microsoft.

Karen Hao	*Empire of AI*	*Chapter 9*	Excerpted by *Andrew M.*

Just as the first era of AI commercialization laid the groundwork for the generative AI era's amassing of data and capitalization of compute, so, too, did it create the foundations for its wide-scale labor exploitation.

Before generative AI, self-driving cars were the biggest source of growth for the data-annotation industry. … As billions of new dollars flooded into the race to create the cars of the future, the demand for data annotation exploded and created a need for alternatives to Mechanical Turk [a then-popular crowdsourcing platform where businesses can post small data-management tasks for which individuals can compete for pay].

But right as this new wave of companies sought to establish themselves, a strange thing happened. Sign-ups on their worker-facing platforms came rushing in from an unexpected country: Venezuela. In the same moment that auto giants began scrambling, money began pumping into self-driving cars, and data-annotation firms began looking for more workers, Venezuela was nose-diving headfirst into the worst peacetime economic crisis globally in fifty years.

Amid the catastrophe, many Venezuelans turned to online platforms for work. By mid-2018, hundreds of thousands had discovered and joined the data-annotation industry, accounting for as much as 75 percent of the workforce for some outsourcing firms. Working on data-annotation platforms became a whole-family activity. ... [In many cases] parents and children often took turns to work on a shared computer; wives reverted to cooking and cleaning to allow their husbands to earn just a little more money by pulling longer hours uninterrupted. The crisis left an indelible mark on the wave of AI-specialized outsourcing firms as they grew up alongside it.

Scouting workers in crisis could become a surefire way to continue driving down the costs of the labor that serves as the lifeblood of the AI industry. Looking back several years later, that's exactly what happened—and what has become one of the most stunning parallels between empires of old and empires of AI. **(205, 206–208)**

Tech colonialism calls on us to rethink what we mean by "postcolonial" and whether or not that term still has substance.

Part of the process of colonialism is not just dominating a group of people but also "cultivating" them in a certain sense (i.e., rendering a piece of land profitable for a market)

By mid-2021, as Venezuelans burned out and left the platform, Scale [one of the most successful data-annotation firms that rose to meet the demands of the self-driving car boom] was scouting and onboarding tens of thousands more workers from other economies that had collapsed during the pandemic. To support its expanding and diversifying client needs, it entered countries with large populations facing financial duress and who could also speak the most economically valuable languages: English, French, Italian, German, Chinese, Japanese, Spanish. It sought French speakers from former French colonies in Africa, an employee who worked on international expansion remembers; it sought Mandarin speakers from places with large populations of Chinese diaspora such as in Southeast Asia.

Scale proceeded to repeat the playbook it had developed in Venezuela again and again. It offered high earnings in each new market to attract workers and throttled those earnings as it settled in. It tinkered with the size of its payouts to taskers through rounds of experimentation that full-time employees, sitting in its now 180,000-square-foot San Francisco headquarters, discussed as optimization and innovation. Workers meanwhile saw their livelihoods decimated with the unpredictable changes. The Scale spokesperson said the company rejected the characterization that it has targeted economies in hardship and purposely cut back earnings. Scale recruits workers based on considerations including geographic and linguistic diversity and 24/7 coverage. "We care deeply about our contributors and any claim to the contrary is false," he said.

One group of eight workers in North Africa said Scale reduced their pay by more than a third in a matter of months. At least one worker was left with negative pending payments, suggesting that he owed Scale money. When the group attempted to organize against the changes, the company threatened to ban anyone engaging in "revolutions and protests." Nearly all who spoke to me were booted off the platform. The Scale spokesperson said the company does not suspend workers for concerns about pay, only violations of Community Guidelines. **(216)**

The legacy of historic colonialism have set the conditions for new paradigms of colonialism to take root and proceed – a kind of crop rotation that exceeds the scope of one lifetime.

neoliberal colonialism as crop rotation

When I wrote the story of [the Kenyan workers] for The Wall Street Journal, OpenAI sought to distance itself from the responsibility of the toll its project exacted. It was Sama that had followed inadequate procedures to protect their workers, OpenAI leadership said; with Sama's pristine reputation before then, OpenAI couldn't have known that the workers were struggling.

But the consistency of workers' experiences across space and time shows that the labor exploitation underpinning the AI industry is systemic. Labor rights scholars and advocates say that that exploitation begins with the AI companies at the top. They take advantage of the outsourcing model in part precisely to keep their dirtiest work out of their own sight and out of sight of customers, and to distance themselves from responsibility while incentivizing the middlemen to outbid one another for contracts by skimping on paying livable wages. Mercy Mutemi, a lawyer who represented Okinyi and his fellow workers in a fight to pass better digital labor protections in Kenya, told me the result is that workers are squeezed twice—once each to pad the profit margins of the middleman and the AI company.

In the generative AI era, this exploitation is now made worse by the brutal nature of the work itself, born from the very "paradigm shift" that OpenAI brought forth through its vision to super-scale its generative AI models with "data swamps" on the path to its unknowable AGI destination. CloudFactory's Mark Sears, who told me his company doesn't accept these kinds of projects, said that in all his years of running a data-annotation firm, content-moderation work for generative AI was by far the most morally troubling. "It's just so unbelievably ugly," he said. **(222–223)**

To Scale AI, Kenya had one advantage that Venezuela did not. The workers speak English, like the chatbots who need them. As self-driving car work largely disappeared from the platform, so did Venezuelans. "They wouldn't use Venezuelans for generative AI work," says a former Scale employee. "That country is relegated to image annotation at best." Scale would soon ban Venezuela from its platform completely, citing "changing customer requirements."

The first time Scale came to Kenya, it had set up physical office spaces for workers to report to and attend trainings. This time was different. It had shuttered those spaces during the pandemic and shifted to entirely remote recruitment and operations—blasting ads on LinkedIn, creating online training courses, and placing people as it had in its other locations in moderated community discussion channels. For workers, it both allowed more flexibility and became much harder to connect with one another, diminishing their chances of organizing for better pay or working conditions like the workers at Sama.

In December 2022, days after ChatGPT's release, [Kenyans working in the remote annotation space] discovered a new type of project under the category "transcription." It wasn't really transcription. All of the projects were asking her to write prompts and example answers for new chatbots from companies now jostling to compete with ChatGPT. [These exercises resembled the RLHF (reinforcement learning from human feedback) training that led to InstructGPT's—one of the precursors to ChatGPT—success. These jobs too, however, would eventually dry up.]

Less than a year later, [Kenyans] would learn the truth. In March 2024, Scale would block Kenya wholesale as a country from Remotasks, just like it did with Venezuela. For Scale, it was part of its housecleaning—a regular reevaluation of whether workers from different countries were really serving the business. Kenya, they decided, along with several other countries including Nigeria and Pakistan, simply had too many workers attempting to scam the platform to earn more money. Such behavior undermined the integrity of the quality Scale delivered to its customers and could risk it losing multimillion-dollar contracts. It simply wasn't worth it.

In a great irony, many of those so-called scams were in fact workers using ChatGPT to generate their answers and speed up their productivity. For white-collar workers in the Global North, such an act, within Silicon Valley's narrative, would be laudatory and, with enough widespread adoption, do wonders for the economy; in the hands of RLHF workers in the Global South, whose very labor props up that narrative, it was a punishable offense. **(229–230, 231, 233)**

Key words/phrases

236-7 (PDF)

"We were young, making salaries relatively standard in the tech industry that placed us nationally in our age group's top 5 percent.

But there was that dissonance. On the way to work, I would pass people shooting up drugs in front of the subway stations, the unhoused peeing on sidewalks just blocks from my office. Meanwhile, our startup's chef, playfully named "the happiness engineer," would cook or cater an abundance of food for our free office lunches. Leftovers often went straight into the trash. If we stayed late, we got free dinner—and were emphatically implored to take an Uber home for safety reasons. It was all too easy for the privileged to grow accustomed to moving through the city in ways that shielded them from seeing the realities of how the other half lived.

The dichotomy encapsulated how the tech industry could profess big, bold visions about changing the world and building a better future while ignoring the very problems at its door. It was a dichotomy that Altman would sometimes comment on in his own way—getting right up to yet never fully acknowledging the utter contradiction of declaring the problem of creating and managing beneficial AGI possible, but San Francisco's housing crisis too tough to tackle."

Problems that don't have clear returns are treated as "too tough to tackle."

A big part of Hao's project is to disabuse readers of the sage-like fantasy-thinking cultivated by Big Tech.

241 (PDF)

"The growing membership in the AI safety community, which knit together EA-backed AI safety with other strains of catastrophic, existential, and risk-focused thinking, swelled Anthropic's ranks just as it restocked OpenAI's Safety clan….The influx of members in AI safety also popularized the community's lexicon more broadly in the AI industry. How fast you think AI will advance and reach major milestones like AGI is your "AI timeline." How likely you think it is that AGI will lead to catastrophic outcomes, meaning the killing off of most of the human population, or existential outcomes, meaning the complete and total extinction of humanity, is your "p(doom)," short for probability of doom. "Hardware overhang," as referenced in OpenAI's 2021 research road map, is another dictionary entry, as is "AI takeoff," the process of AGI improving to the point of superintelligence and thus capable enough to outwit humanity. "Acceleration risk" refers to "the risk of triggering a heightened competition between companies or countries that leads to a potentially dangerous acceleration of AI advancement and a shortened AI timeline."

"But for a movement that professed independent thinking, EA was swiftly accelerating in the opposite direction. People attracted to its premise were quickly indoctrinated into a broader set of dogmas, propelled by the promise of more opportunities and resources, and an insular social network that played fast and loose with personal and professional boundaries. Within Silicon Valley in particular, EA people largely worked only with other EA people; they largely lived, partied, dated, and slept only with other EA people. Mixed with the tech industry's deep-rooted sexism and the Bay Area's long-standing polyamorous subcultures, its cultlike fervor, manifested in the worst way, could turn into a toxic cauldron of sex, money, and power; it was leading EA to be plagued by growing allegations of sexual harassment and abuse."

246-7 (PDF)

"After the experience of firefighting text-based child sex abuse with AI Dungeon, of particular concern was the possibility of DALL-E 2 being used to manipulate real or create synthetic child sexual abuse material, or CSAM. As with each GPT model, the training data for each subsequent DALL-E model was growing more and more polluted. For DALL-E 2, the research team had signed a licensing deal with stock photo platform Shutterstock and done a massive scrape of Twitter to add to its existing collection of 250 million images. The Twitter dataset in particular was riddled with pornographic content. Several employees made a significant effort to check for and cull any CSAM.

But after some discussion, the employees left in other types of sexual images, in part because they felt such content was part of the human experience. Keeping such photos in the

training data, however, meant the model would still be able to produce synthetic CSAM. In the same way DALL-E could generate an avocado armchair having only ever seen avocados and armchairs, DALL-E 2 and DALL-E 3 could do the same thing with children and porn for child pornography, a capability known as "compositional generation….Later, during the development of DALL-E 3, when the data imperative had grown even larger, the research team decided that sexual images were no longer just a "nice to have" but a "need to have." The share of pornographic images on the internet was so large that removing them shrank the training dataset enough to notably degrade the model's performance. In particular, it made the model worse at generating faces of women and people of color due to the same discovery that Deborah Raji made as a Clarifai intern: A significant share of the online content depicting both groups is sexually explicit. For the same reasons, the researchers left in some other kinds of disturbing images."

250-1 (PDF)
"In March 2022, OpenAI released DALL-E 2 via the Labs web app to overwhelming public enthusiasm. As people gushed over and grappled with the model's capabilities, to a degree that exceeded many employees' expectations, the web app went viral across social media, producing a plethora of wild, wacky, and surreal AI-generated art in its wake. It was a GPT-3 moment but better. Instead of engaging with only a small pool of technical developers, the company was tapping into a much broader and more global base of consumers. In real time, it could also respond to user feedback with instantaneous changes to the Labs web app. "This is intoxicating," Fraser Kelton would remember of the experience in a podcast."

259-60 (PDF)
"As they raced to set up product policies and monitoring and enforcement infrastructure, members of the budding team felt a constant sense of uncertainty about how trust and safety for an AI company should differ from a search or social media company and whether their preparations for GPT-4 were adequate. Trust and safety was typically focused on preventing a predictable slate of internet abuses, like fraud, cybercrime, and election interference. But wrapped up in the confusion was how their work related to that of the AI safety people within Leike's alignment and Brundage's policy research teams who often discussed unknowable, catastrophic doom. OpenAI labeled all of them as "safety" teams, but they seemed to be speaking fundamentally different languages. Although the vocabulary of existential AI safety had, with its popularization through EA, become common parlance in the field, employees coming from traditional tech company backgrounds had never heard of AI timelines or quantified their p(doom). Even shared words between the two groups like *risk* and *harm* seemed laced with different connotations."

This section echoes the start of the chap. in which she talks about the relationsh-ip in tech between elitism and insularity

265 (PDF)
"That September, the technical leadership held an off-site at the Tenaya Lodge, a remote luxury resort nestled in the lush folds of the Sierra Nevada. The property, furnished with stunning interiors and multiple pools and restaurants, sat only two miles away from the millions of acres of pristine wilderness in Yosemite National Park. On the first night, everyone gathered around a firepit on the rear patio of the hotel. Senior scientists, dressed in bathrobes, flanked the fire in a semicircle.

Then Sutskever emerged. In the pit, he had placed a wooden effigy that he'd commissioned from a local artist, and began a dramatic performance. This effigy, he explained, represented a good, aligned AGI that OpenAI had built, only to discover it was actually lying and deceitful. OpenAI's duty, he said, was to destroy it. Only a few yards away, several redwoods stood like ancient witnesses in the darkness. Sutskever doused the effigy in lighter fluid and lit it on fire."

Key words/phrases

Karen Hao	*Empire of AI*	*Chapter #11*	Excerpted by *Zach W.*

258; 260–261

The company wouldn't wait to ready GPT-4 into a chatbot; it would release Schulman's chat-enabled GPT-3.5 model with the Superassistant team's brand-new chat interface in two weeks, right after Thanksgiving. The Superassistant team instantly pivoted, pulling in several other members as they sprinted to integrate everything and build out the remaining features. To the rest of the company, leadership framed the effort carefully. ChatGPT—the name they settled on—would not in fact be a product launch but a "low-key research preview," just like DALL-E 2. In the same way, it wouldn't be monetized but "get the data flywheel going"—in other words, amass more data from people using it—which would help improve GPT-4 and the Superassistant product.

…

In the immediate aftermath [of ChatGPT's launch], the whole company was firefighting. OpenAI's servers crashed repeatedly as the infrastructure team struggled to scale up its capacity as fast as possible, in the most compressed timeline in the history of Silicon Valley. The team cannibalized some of the Research division's compute to support ChatGPT's growth and still didn't have enough to keep the app up and running. The trust and safety team, numbering just over a dozen people, scrambled to understand and catch bad behavior among the floods of new users, heavily handicapped by spotty monitoring. It had struggled to hire the engineers needed to implement its limited reactive enforcement plan and was still in the middle of building the necessary systems. ChaptGPT derailed the project. All efforts to finish new tooling halted as engineering resources were redirected to stabilize what already existed. When the servers crashed, so did the platform for monitoring traffic, grinding the ability to do any scaled enforcement to a complete halt.

262

With every team stretched dangerously thin, managers begged Altman for more head count. There was no shortage of candidates. After ChatGPT, the number of job applicants clamoring to join the rocket ship had rapidly multiplied. But Altman worried about what would happen to company culture and mission alignment if the company scaled up its staff too quickly. He believed firmly in maintaining a small staff and high talent density. "We are now in a position where it's tempting to let the organization grow extremely large," he had written in his 2020 vision memo, in reference to Microsoft's investment. "We should try very hard to resist this—what has worked for us so far is being small, focused, high-trust, low-bullshit, and intense.

… "The overhead of too many people and too much bureaucracy can easily kill great ideas or result in sclerosis. Unlike the big-iron engineering projects of the past, we could fulfill our mission with a surprisingly small number of great people."

… Over several weeks, as the discussions continued, the executive team finally compromised on a number somewhere in the middle, between two hundred fifty and three hundred. The cap didn't hold. By summer, there were as many as thirty, even fifty, people joining OpenAI each week, including more recruiters to scale hiring even faster. By fall, the company had blown well past its own self-imposed quota.

262–263

The sudden growth spurt indeed changed company culture. A recruiter wrote a manifesto about how the pressure to hire so quickly was forcing his team to lower the quality bar for talent. "If you want to build Meta, you're doing a great job," he said in a pointed jab at Altman, alluding to the very fears that the CEO had warned about of the company rapidly diluting its talent density and mission orientation, while increasing its bureaucracy. The rapid expansion was also leading to an uptick in firings. During his onboarding, one manager was told to swiftly document and report any underperforming

The phrase "low-key research preview" suggests the model didn't have to go through rigorous testing.

The decision to release early also demonstrates how OpenAI skipped creating a product and rather rolled out the product development chat bot 'as' the product.

It seems like OpenAI has become a bit of a runaway freight train

members of his team, only to be let go himself sometime later. Terminations were rarely communicated to the rest of the company. People routinely discovered that colleagues had been fired only by noticing when a Slack account grayed out from being deactivated. They began calling it "getting disappeared."

… To new hires, fully brought into the idea that they were joining a fast-moving, money-making startup, the tumultuousness felt like a particularly chaotic, at times brutal, manifestation of standard corporate problems: poor management, confusing priorities, the coldhearted ruthlessness of a capitalistic company willing to treat its employees as disposable. "There was a huge lack of psychological safety," says a former employee who joined during this era. "It is like the opposite of 'a company as a family'—which is fair, you know, it is a company." Many people coming aboard were simply holding on for dear life until their one-year mark to get access to the first share of their equity. One significant upside: They still felt their colleagues were among the highest caliber in the tech industry, which, combined with the seemingly boundless resources and unparalleled global impact, could spark a feeling of magic difficult to find in the rest of the industry when things were actually aligned. "I would say OpenAI is one of the best places I've ever worked but also probably one of the worst," the former employee says.

… For some employees who remembered the scrappy early days of OpenAI as a tight-knit, mission-driven nonprofit, its dramatic transformation into a big, faceless corporation was far more shocking and emotional. Gone was the organization as they'd known it; in its place was something unrecognizable. "OpenAI is Burning Man," Rob Mallery, a former recruiter, says, referring to how the desert art festival scaled to the point that it lost touch with its original spirit. "I know it meant a lot more to the people who were there at the beginning than it does to everyone now."

265

Many [Microsoft] executives were no longer talking merely about beating Google. Where they had once responded to OpenAI's strange talk about AGI and highfalutin language about its power to invent the future with polite skepticism, they were now believers. AI, AGI, generative AI—whatever you wanted to call it—this technology *was* the future, and Microsoft was shepherding it hand in hand with OpenAI. The more Microsoft believed, the more the company reoriented its rhetoric and strategy. "I saw the incentives at Microsoft push more and more toward a narrow conception of the future," the former employee says. "I saw the technology become narrowed into something propped up by narrative rather than reality."

… Nadella implemented a new strategy for distributing Microsoft's computing resources. He shifted GPUs away from Microsoft's research teams to support OpenAI. The company also consolidated all of its GPUs into one pool for better supporting generative AI workloads. "The typical Microsoft employee had no fucking clue what OpenAI was before January last year," one Microsoft employee remembers. Now they were receiving urgent directives from their superiors about finding ways to intersect their work with OpenAI technologies.

Key words/phrases **"Low-key research preview" ; scale ; psychological safety ; Arrakis ; Stargate**

Karen Hao	Empire of AI	Chapter 12	Excerpted by Andrew M.

Chile is the world's largest producer of copper, accounting for a quarter of the global supply. Since the beginning, copper mining has reshaped not just the land but the societies that rely on it. Sometimes the effects are visible: Chuquicamata today is the largest open-pit copper mine in the world.

Lithium is a more recent discovery there, stumbled upon by an American company in the 1960s as it searched for the water it needed for copper mining. When it drilled into the *salares* [salt flats], it found high concentrations of lithium floating in an oily brine beneath the surface, opening up a new front of extraction and accelerating the depletion of more ecosystems.

Over the years, the *Atacameños* [the term that names the multitude of tribes in northern Chile] have heard many narratives used to justify all of this extraction. In 2022, as the European Union set new policies around the energy transition and the demand for lithium skyrocketed, both companies and politicians in Chile and the rest of the world lauded the importance of the country's mining industry in propelling forward a better future. Indigenous communities watching their land and their communities get ripped apart asked: A better future for whom? "Local people never have the ability to think about their own destiny outside the forces of economics and international politics."

Now the same narratives are being recycled with generative AI. The accelerated copper and lithium extraction to build [data centers]—and to build the power plants and thousands more miles of power lines to support them—is, in Silicon Valley's account, also ushering in a better and brighter future. To block that extraction is thus to block fundamental progress for humanity. But it is not the mining that Indigenous communities resist. "Our ancestors were miners," says [one tribal member]. They were the ones who discovered the copper in the first place. The problem, she says, is the scale.

That scale has consumed everything. It has made the north and the rest of Chile completely dependent on the industry and not allowed for the emergence of other economies. It has choked off the country's—and the rest of the world's—ability to imagine different paths. **(293, 294–295)**

Tech's narratives of process overwhelm what alternatives are conceivable.

[In 2012, Google announced that it would build its first data center in Latin America in Quilicura, a commune in Chile's capital, Santiago.] On Google's web page, the company proudly presents the data center, which became operational in January 2015, as one of the most efficient and environmentally friendly on the continent. At the time, no one in Quilicura paid attention to the project; certainly no one in the better-off epicenter of Santiago was paying attention to Quilicura. Neighborhood residents who passed by the data center every day on their bus route to work assumed that it was a factory producing beer or food and providing jobs to the local community.

Google's data center … did not provide many jobs beyond its initial construction. A job posting from 2024 for a mechanical technician, one of the few long-term positions available, was advertised on Google's job board only in English; the posting stated that Google wouldn't consider résumés submitted in any other language. The data center—as activists point out—did not provide much other benefit to the local community either. Nearby, public schools still lack good internet or devices for students to access it.

[In] July 2019, as Google began the paperwork for its second data center in Chile, a group of residents was watching. The company had chosen Cerrillos for its new location, another working-class municipality bordering Santiago. ... That summer, as Google filed a report with Chile's environmental agency for approval of its data center—a largely rubber stamp process—MOSACAT, a water activist group, began combing through all 347 pages of the filing. Buried in its depths, Google said that its data center planned to use an estimated 169 liters of fresh drinking water per second to cool its servers. In other words, the data center could use more than one thousand times the amount of water consumed by the entire population of Cerrillos.

[After MOSACAT alerted local officials of Google's estimates, the company] sent two engineers and a lawyer to Cerrillos to present to the community. … The Google engineers were gringos, MOSACAT remembers, tall and able to speak only English. They gave a highly technical presentation, and Google's lawyer doubled as a translator. During the discussions, MOSACAT

What Google is doing effectively amounts to greenwashing by emphasizing the environmentally conscious aspects of their industry while conveniently failing to mention that astronomical amount of fresh water needed to cool their systems.

members who spoke English say they could hear the lawyer mistranslating their words. At another point, the Google representatives sought to assuage the community by offering to plant an urban forest just like the one the tech giant had given to Quilicura. The show left MOSACAT and the other groups unimpressed: Google wasn't here to truly engage with and hear what the community wanted. "They came to intimidate us," says a MOSACAT member Alejandra Salinas, who also serves as a council member for Cerrillos's neighboring municipality, Maipú. "Think about it. They come offering us trees while drying out our earth." **(298–299, 300, 301–302)**

In December 2019, MOSACAT won; the referendum to build the data center was rejected with a slim majority of the vote. The following year, the government joined MOSACAT in filing a lawsuit against the project with an environmental court.

In the meeting with Google, its representatives ultimately put on a friendly face, MOSACAT says, presenting a greater willingness to negotiate once it became clear that some residents could understand them. ... [The] risk of pushing back as a Global South country is always that a Silicon Valley company will pick up and take its money somewhere else. As the project continued to stall and Google's desire for more compute intensified, the company announced that it would shift its next planned data center [to Uruguay.]

In Uruguay, more than 80 percent of the country's fresh water goes to industry instead of human consumption—most notably, cash crop agriculture. These include industrial farms for soybeans and rice, and for trees that feed into paper production. Most such farms are run not by local companies but by multinationals that export what they grow. … Their activities deplete the nutrients in the soil, making it more difficult to grow actual food, and pollute the country's water streams.

As with Chile, the foreign multinationals still exist above locals in the political pecking order. During the drought, industry continued to use water unabated, drawing what it needed directly from the main river … that also feeds Montevideo's public water system. Two decades ago, after significant environmental activism, Uruguay became the first country in the world to recognize water as a human right in its constitution. Now, in a bitter irony, it was the drinking water rather than water for industry that saw the most severe cutbacks during the shortage.

[Because access to water is a human right in Uruguay, local activists were unusually successful in compelling officials to release sealed documents that detailed Google's projected water consumption.] The environmental ministry revealed [Google] planned to use two million gallons of water a day directly from the drinking water supply, equivalent to the daily water consumption of fifty-five thousand people. [In response] thousands of Uruguayans took to the streets to protest Google and all of the other industries that had led the government to squander the country's precious freshwater resources.

[As one activist put it, these data centers are] "extractivist projects that come to the Global South to use cheap water, tax-free land, and very poorly paid jobs. And then they don't contribute to our country; they don't improve our internet access." **(305, 306)**

Judging from the repeated mention of internet connectivity by South Americans suggests that one of the promises put forward by tech companies is that they're going to invest in the internet connectivity of South America.

Hao is not critical enough to the right to water here because she sees some promise in rights-based approach to challenging tech companies.

As data center developers go, Microsoft was a late bloomer. Before its investments in OpenAI, Google was well ahead in the number of facilities it was constructing around the world. But with the sudden explosion of demand for more computing infrastructure to support its AI ambitions, Microsoft adopted Google's playbook and followed it into the same regions.

The Chile that Microsoft entered was different from the Chile that had greeted Google. By 2022, more than two years had passed since the *Estallido Social* [a massive popular uprising between 2019 and 2021 that led to constitutional reform in Chile], which had left an indelible mark on the country's politics.

Upon Microsoft's entrance into Quilicura, [a new vanguard of activists, Resistencia Socioambiental Quilicura,] did what MOSACAT and Daniel Pena had before them: They began to pore over whatever materials they could find that Microsoft had made available and to extensively research the project.

Microsoft projected that it would need a significantly lower amount of water than Google had in Cerrillos. Even still, [one *Resistencia* activist worried]. The drought in Chile had only gotten worse and was expected to last until 2040. Quilicura's wetlands were suffering acutely, on top of industrial encroachment, from accelerating desertification. On its website, Microsoft boasted about new cutting-edge innovations in data center cooling systems that would mean its facilities didn't need to use water. If Microsoft had the capability to build waterless data centers, why wasn't it doing so in Quilicura?

Microsoft would subsequently explain that the innovations it had advertised were still under development and only being piloted in a place in the US. "Then why do they promise these things" on their website?"

[Today, three years into his four-year term, Chile's new president] is under pressure to get "quick wins" for the economy—even more so as a young, left-wing president. That means pressure to expand the mining industry, pressure to see through the arrival of twenty-eight new data centers.

the coalition of activists in northern Chile and Santiago with researchers domestic and abroad has clearly made a mark. As part of the data center plan, the Ministry of Science, which oversees its creation, has for the first time formed a committee of activists to consult with regularly as part of the drafting process, and invited [Resistencia Socioambiental Quilicura] and members of MOSACAT to join them. The Ministry of Science is also overseeing an AI bill that articulates how Chile wants to approach AI development, application, and regulation. Whereas before the ministry's discussions cast AI as a universal positive, the tone has since shifted to acknowledge the social and environmental costs of the technology. **(307–310, 311–312)**

The dynamics at play in the Estallido Social are also present and iterate differently across every single populist movements across the globe.

Key words/phrases

Community Quote Sheet

Metaphor Key: ★ (recurrent); ⊥ (contradictory); ƨ (abstruse)

Karen Hao	*Empire of AI*	*Chapter 13*	Excerpted by *Tracy*

Pg 305 "Amid the climate of frustration and fear in Washington, a policy white paper echoing Altman's recommendations arrived two months after his hearing in July 2023. Written by a consortium of researchers, including from OpenAI's Safety clan, Microsoft, and Google DeepMind as well as more than a dozen think tanks, many tied to the Doomer community, it pushed once again for a new licensing regime for AI models using compute thresholds, and the development of AI safety evaluations for dangerous capabilities including the ability to manipulate and persuade and the creation of novel biological weapon recipes. The fifty-one page document also gave a name to the category of models that needed this government intervention: "frontier" AI models. Per the authors, frontier models —models that might exhibit these dangerous capabilities — did not yet exists, but by scaling existing models from companies like OpenAI and Anthropic with ever more compute, they could arrive suddenly and unpredictably at any moment….It was a rare issue in which the interests of Doomers, Boomers, and profit-motivated corporates aligned: Keeping frontier models front and center in government regulatory discussions was ideologically imperative to Doomers and convenient to Boomers and corporates for shifting attention away from regulating existing AI models and their problems."

These quotes linger on the idea of safety in Hao's book, spefically how OpenAI keeps attention on far off threats of AGI to deflect attention away regulating AI in the here and now.

Pg 306 "But Hooker and many other researchers, including Deborah Raji, disagree with the compute-threshold approach for regulating models. While scale can lead to more advanced capabilities, the inverse is not true: Advanced capabilities do not require scale. A deep learning model trained only on high-quality biological data, for example, can be a very powerful generator of biological recipes at very small scale. Through distillation, one of the techniques that OpenAI referenced in its 2021 research road map, large models can also be transformed into small models with similar capabilities."

What makes OpenAI so compelling is that the company can make a generalized sales pitch without specifying precisely what it is that they're selling.

308 "As Commerce consulted various experts on its proposal to clamp down on model weights, news of its deliberations, which it would announce in early 2024 with a public request for comment, cleaved the AI development community into two. This clash was about Closed versus Open, techno-nationalism versus borderless science. In addition to the Frontier Model Forum participants and broader Doomer community, the Closed side quickly won over the US national security and intelligence apparatus. Facing off against them was Meta, open-source AI developers, startups, civil society groups, and independent academics."

Key words/phrases	safety

Community Quote Sheet

Metaphor Key: ★ (recurrent); ⊥ (contradictory); ƨ (abstruse)

Karen Hao	Empire of AI	Chapter 13	Excerpted by Tracy

Pg 319 "In 2023, they [OpenAI researchers] believed those threats now included targeted attacks and rogue AI itself. On Slack, the security team posted a draft of its threat model for OpenAI, significantly matured from the days when executives had debated how much to heighten the security of the organization. The draft included three categories of threats: foreign state actors, competitors, and ideologically motivated people.

In the third category, the draft linked to an article by Eliezer Yudkowsky, an extreme Doomer and leader in the AI safety community who had coined and popularized the phrase *friendly AI* to refer to well-aligned systems…He proposed a plan to enforce the halting of AI advancement by shutting down all large GPU clusters, tracking sales of GPUs, and if necessary, targeting a "rogue data center' with air strikes. OpenAI's threat model draft linked to this piece as an example of an ideologue advocating for violence.

A small faction of Open AI employees who were fans of Yudkowsky's less violent opinions found the fact that the draft singled him out but didn't mention unfriendly AI itself deeply frustrating. After getting feedback, the security team added a fourth threat category: misaligned AGI systems."

Key words/phrases

Karen Hao	*Empire of AI*	*Chapter #14*	Excerpted by *Zach W.*

Key words/phrases	**Microcosm; financial independence; "mental health challenges"; atonement; shadow ban; abundance**

Questions	1. **Why is readily abundant material aid withheld until a hypothetical far-off future?** 2. **Why does "AI" exacerbate the desperate conditions of people in crisis?** 3. **Why do interpersonal relations appear to parallel commercial relations?**

330–331	After her dad's death in 2018, Annie [Altman] was living in LA. She was working part time as a writing assistant and then at a marijuana dispensary. Annie says she inherited around $100,000 from her dad's life insurance. She threw herself deeper into pursuing her love of art and performance as a serious career: She took comedy classes; she went to open mics; she started a podcast; she rewrote the first draft of her book into a one-woman show she renamed The HumAnnie. By mid 2019, the funds were depleting, most of it spent, she says, on her various artistic investments, her high rent, her out-of-pocket health insurance, and her accumulating medical expenses—medical imaging, physical therapy, psychotherapy, Lyfts and Ubers to her appointments. … For Annie, it was what happened next that spelled the beginning of the end of her relationship with her family. In May 2019, as her chronic pain continued, she learned that her dad had left her his 401(k). She quit the dispensary and drew up a six-month plan to use the extra financial runway—a little over forty thousand dollars—to tend to her health and get her creative endeavors off the ground enough to hopefully generate a sustainable source of income. But soon after, her mother [Connie Gibstine], who retained authority over her dad's retirement funds as the surviving spouse, notified Annie that she would not in fact be receiving the money. The best tax strategy, Gibstine wrote to Annie in an email, was for Gibstine to keep the funds in her name in a tax-deferred account that would pass to Annie via a trust and to which she would gain access at age fifty-nine and a half. "We all want what is best for you, and **we believe that what is best for you (and for everyone) is financial independence, which brings long-term satisfaction, personal growth and security**," Gibstine said. "For this reason, we want to clarify that we will not financially support you if/when your current inheritance from Dad runs out." The email was signed "Mom, Sam, Max, Jack." … Annie was in a fragile state. Her therapist's notes from the same period say that the move her family made with the hope of getting her back on her feet instead worsened her condition. In December 2019, her bank account slid into the negative. Sam had secured Microsoft's first $1 billion investment into Open AI earlier that year and the company was full speed ahead in training GPT-3. Scared and alone, Annie logged on to an escort service called SeekingArrangement and showed her breasts to a man over a video call for enough money to bring her account out of the red.	

	… Annie says she had never before asked her mother or brothers for financial support. She had not assumed she could rely on their wealth, but neither had she believed they would leave her without a safety net should she truly be in an emergency.	
332–333	Over the next three years,…Annie's life bottomed out. She faced housing insecurity, food insecurity, health insecurity; she turned to virtual and physical sex work to pay the bills. In their public statement, the family said they tried "in many ways to support Annie and help her find stability." In Annie's retelling, one of those extensions of support came in the spring and summer of 2021, when Sam sought to reconnect with her. They had three phone calls, she remembers, during which he told her how much he loved her and offered to buy her a house. Annie and the family describe that offer differently. Annie says the offer was not for her to own the house but for her to live in it, an arrangement she understood was meant to prevent her from selling the property and which she worried could be another way through which Sam and their mother could impose their views on her health and career decisions. Gibstine said in her statement to me that the family offered Annie home ownership but did not respond to my requests for corroborating documentation. In the family's public statement, they said they offered to buy Annie a house through a trust, so that she could have a place to live without the ability to sell it immediately. In the end, Sam and Annie reached a different agreement: He would pay her rent for a year directly to her landlord. In the summer of 2022, when she didn't reestablish contact, the payments stopped.	
333–334	Despite all the leaps and bounds in AI capabilities, none of them helped to alleviate any part of Annie's desperation. If anything, AI may have served to entrap her further. She hadn't wanted to turn to sex work. It was, she says, a "plan Z." When her chronic pain intensified and made even her work trade too difficult, she had first attempted digital means of monetizing her art. She continued her podcast and maintained an Etsy store and Patreon account, but they didn't earn enough to even cover her phone bill. A strange thing was happening, which she documented in screenshots over time: She was getting little to no exposure across all of her social media. Sometimes she noticed chunks of the reviews on her podcast in the Apple app mysteriously disappearing, which limited its discovery. At least twice, on both her Instagram and YouTube, she would accumulate views and then inexplicably lose the,. A former Facebook data scientist and two tech and sex work experts say it's possible that Annie's very first SeekingArrangement account in December 2019 could have limited her online traction, based on the nature of how tech platforms track and shadow ban sex workers through automated systems by tagging their devices, emails, bank accounts, or other informations that ties together their online presence, even for profiles completely unrelated to their sex work. Seeing no other path forward, Annie went back to SeekingArrangement, entangling her online presence and access to economic opportunities even more in a web of algorithmic moderation.	
333; 339–340	Annie's story deepens the dueling portraits that people paint of Sam. He is at once generous and self-serving, agreeable and threatening, a benefactor for so many people and the source of great personal pain for others. Someone who projects sincerity and altruism in public but reveals a more complicated calculus through	

his behaviors behind closed doors. Someone who can give and take away, leaving many with an impression that they are part of a larger game of chess for which only he can see the full board, and the end game is to preserve his power as king.

…

Annie's story also complicates the grand narrative that Sam and other OpenAI executives have painted of AI ushering in a world of abundance. Altman has said that he expects AI to end poverty. Brockman has repeated, through his stories about his friend and his wife, Anna, that AGI will dramatically improve healthcare. Sutskever has said that it will lead to wildly effective, dirt-cheap psychotherapy. And yet, against the reality of the lives of the workers in Kenya, activists in Chile, and Altman's own sister's experience bearing the brunt of all of these problems, those dreams ring hollow.

…

Since resolving to tell her story, Annie has faced the same gulf of power that I have watched data workers and data center activists wrestle with. Her life has revolved around combing through and gathering as much documentation as possible to get anyone to listen to her. At times she has been consumed by a sinking feeling that no matter how much she speaks up, the world is somehow in a conspiracy against her. It's the same loss of agency and anger I've seen etched on the faces of people globally when they throw so much of the little they have at challenging the empires' narratives, and then watch as the people they are up against wield the kind of power that can deploy billions of dollars in capital, construct vast infrastructure, hire and fire tens of thousands of contractors, and, with a few soft-spoken words—at an event, to Congress, to heads of state, to journalists—smooth over the murmurs of protest in the way of their will.

Altman's behavior had progressively worsened after ChatGPT had propelled him into megastardom, intensifying both the spotlight and scrutiny and exploding his calendar with an overwhelming travel schedule. Before, he was generally energized; now he was often exhausted. And he was cracking under that pressure, his anxiety reaching new heights and fueling his patterns of destructive behavior. He was doing what he'd always done, agreeing with everyone to their face, and now, with increasing frequency, badmouthing them behind their backs. It was creating greater confusion and conflict across the company than ever before, with team leads mimicking his bad form and pitting their reports against each other. This was corroding enough as it was. But faced with mounting competition externally, Altman was also pushing the company to deploy faster and faster and attempting to skirt some of its established release processes for expediency, sometimes through dishonesty. Recently, he had told [OpenAI's CTO, Mira Murati,] he thought that OpenAI's legal team had cleared GPT-4 Turbo for skipping DSB review. But when Murati checked in with Jason Kwon, who oversaw the legal team, Kwon had no idea how Altman had gotten that impression.

In the summer, Murati had attempted to give Altman detailed feedback on the accelerating issues, hoping it would prompt self-reflection and change. Instead, he had iced her out, and it had taken weeks for her to thaw the relationship, including by assuring him that she had not shared that feedback with anyone else. She had seen him do something similar with other executives: If they disagreed with or challenged him, he could quickly cut them out of key decision-making processes or begin to undermine their credibility. It was subtle and contained enough, out of sight of employees, that it had taken her some years to realize the full extent of it. But inevitably, different executives had each had their turn bearing the brunt of this treatment. Over time, the cumulative impact of his actions had taken its toll on the highest levels of the organization.

[Out of concern for Altman's behavior, Murati eventually reached out to one of OpenAI's board members, Helen Toner, to communicate to her in confidence her belief that the board needs to appoint someone to oversee safety compliance who was not professionally or financially entangled with Altman, elaborating:] Altman was an incredibly anxious person. And when he felt anxious, he had dumb ideas, particularly when enabled by Brockman. Altman's anxiety also fed into toxic behaviors that always followed the same playbook: To anyone resisting his decisions, he would say whatever he thought they wanted to hear to win their support; then, when he lost patience waiting and believed they would continue to go against him, he would undermine their credibility until they got out of the way. It was subtle but pervasive. **(358–359, 362)**

Altman is good at mirroring the desires of others, but he is not skilled at navigating his team through difficult challenges, which is only becoming more pronounced as OpenAI scales.

Is Sam Altman an effective leader or just a skilled courtier?

[Independent of Murati, Ilya Sutskever, OpenAI's then-chief scientist and board member, also reached out to Toner to raise his concerns about his suspicions about Altman's manipulative behavior, which were exacerbated by the surfacing of Annie Altman's sexual abuse allegations against her brother. Sutskever stressed to Toner that whatever actions the board took, simply holding Altman to new performance expectations or increasing the size of the board would be insufficient to the systemic concerns that Altman's leadership posed for the company. Shortly following these admissions by Murati and Sutskever, Toner was contacted by Altman.]

Toner wasn't sure what Altman wanted to talk about. He had texted her earlier that day, October 25, asking if she had time to chat today or tomorrow. Two days before, on October 23, she had spoken again to Sutskever, who seemed far more open after hearing from Murati that she and Toner had also spoken. He made his concerns more explicit than ever before. "I don't think Sam is the guy who should have the finger on the button for AGI," he'd said, and noted the "tremendous opportunity" that had befallen the board to do something about it. He'd then suggested a path forward: replace Altman with Murati as an interim CEO.

[Prior to Sam's text, Toner had met with two other OpenAI board members to discuss Altman.]

Toner wondered whether Altman had somehow caught wind of these meetings and wanted to put an end to the discussions. But now over a call, he was talking about something completely different—concerns he had over a research paper she had published in her day job at CSET.

Is Hao here trying to draw attention to how women in this company are routinely saddled with cleaning up the messes of bad male actors?

Toner had published three papers that week. He singled out the one that had been the most dense and academic. It was about a political science idea called "costly signals," referring to the challenges that state and private actors face when signaling to the public about their intentions with AI regulation and development. She was the third coauthor, and references to OpenAI were buried on pages 28–30 of a sixty-five-page document.

Altman told Toner the paper had been flagged by someone external in an email to OpenAI only a few hours after its release. He was worried that it could look bad for a board member to criticize OpenAI while it was under regulatory scrutiny, including from a July 2023 FTC probe over the company's data, training, and security practices as well as its models' hallucinations that may have reputationally harmed consumers. ... She admitted that she hadn't taken a look with fresh eyes in the context of the new political environment.

The call lasted fifteen minutes. Altman's voice had been mild- mannered throughout. They both agreed at its conclusion that she would email the rest of the board to flag the paper and explain what had happened. In the email, she struck a conciliatory tone. She apologized for two mistakes: believing that the paper wouldn't draw anyone's attention and not reviewing it more closely because of it. [Following Toner's email, she did not hear a response from other board members and therefore believed that was the end of it.] **(369, 370–371)**

Shortly after Altman's call with Toner, Altman had sent out an email to some people within the company saying that he had spoken with [Toner] about her paper and strongly disagreed with her about its consequences. "I did not feel we're on the same page on the damage of all this," he wrote in the email. "Any amount of criticism from a board member carries a lot of weight." To Sutskever, Altman had said more directly that Toner needed to go as a board member and that McCauley [another member of the board] had agreed with him. This felt off to Sutskever.

[Skeptical of Altman's claim, Sutskever called McCauley to confirm whether she had said that Toner needed to go.] On the other end of the line, McCauley seemed dumbfounded. She had definitely not said that, she responded. Altman had indeed called her late on October 24 to talk about the paper. He'd mentioned that it had included a section he thought was critical of OpenAI and that D'Angelo hadn't felt it was a fireable offense but also that Toner shouldn't have written it. McCauley had then told Altman that she hadn't seen the paper and suggested he have a conversation directly with Toner. There was no way she could have said anything to suggest pushing Toner off the board, McCauley said.

It would seem like a coincidence: In the middle of talking about Altman's "specific untruths," here was yet another example playing out in real time of exactly that. Altman would have even gotten away with it had Sutskever not already had reason to reach out to McCauley. Altman knew that the two typically never talked. But to Sutskever, the frequency with which Altman was lying and maneuvering had made it only a matter of time before something like this happened.

After hanging up with McCauley, Sutskever called Toner back. It was time, they agreed, for the three independent board members to talk. **(371–372)**

As a case study, this book is about the interpersonal dimension of corporate capitalism

Key words/phrases

By Saturday, November 11, the independent directors had made their decision [on whether to remove Altman as CEO]. As Sutskever suggested, they would remove Altman and install Murati. … On the night of Thursday, November 16, all four [independent directors] video called Murati. She picked up the call [and with an air of confidence accepted to act as interim CEO.]

Within hours of the public announcement on Friday, November 17, things had gone significantly south for the independent directors. After what had seemed like an initial period of calm and stability, including Murati having a productive conversation with Microsoft, she had suddenly called them with new stress. Altman and Brockman were telling everyone that Altman's removal had been a coup by Sutskever, she said. Combined with Sutskever's ineffectual communication during the employee all-hands, key stakeholders were beginning to turn on the decision.

Shortly thereafter, as the independent directors confronted a hostile leadership over a video call, they realized just how bad the sentiment was. ... [Even] those in the room who they knew … had similar or other reservations about Altman's leadership were remaining silent. As the night wore on and the hostility mounted, the independent directors' two most important allies … were beginning to veer off in a different direction.

[Sutskever, fearing a wave of resignations would lead to OpenAI's dissolution, began pleading with his fellow directors to reroute course. For her part, Murati began acting differently–while exuding confidence in private, she regularly came off in public as reluctant to throw her own weight behind the board's decision to fire Altman, feigning an innocent confusion about the whole situation.]

No longer certain about whether they could rely on Murati, the three independent directors pushed ahead with searching for a new interim CEO or new board members. D'Angelo, in particular, the board's only Silicon Valley insider, made dozens of calls through Saturday and Sunday, putting out as many feelers as possible to his sprawling network. At one point, the directors called Dario Amodei about the interim chief position. Amodei wasn't interested.

On Sunday, D'Angelo finally found someone to take the board up on one of its offers: Emmett Shear, the cofounder of Twitch, appeared willing, and cooperative, to temporarily take over the company. But soon enough he, too, began to deviate for seemingly inexplicable reasons. *The Wall Street Journal* would later report that Shear, who had been YC batchmates with Altman, was also a friend and mentor of YC alumnus Airbnb cofounder Brian Chesky. All weekend, Chesky, among Altman's most trusted friends, had worked the phone lines along with Reid Hoffman, a Microsoft board member, to calm investor nerves, align Microsoft's messaging, and mount the pressure further. Chesky quickly reached Shear, who then sided with Altman.

By late Sunday night, after reassurances from Chesky and Hoffman, Microsoft had also thrown its weight behind Altman, with Nadella announcing that Altman and Brockman were joining to lead a new advanced AI research division. Other AI labs were circling OpenAI like vultures, intent on poaching away their share of talent in the carnage. It became clear to Murati that with the board unable to legitimize their decision, the dynamics now threatened to leave behind a shell of a company. She fell on the side of her fellow leaders. To salvage the situation, Murati wanted Altman back. **(376–379)**

After The Blip [the name used by employees to name the attempted ouster of Altman], Sutskever never returned to the office. [And he would announce his resignation one month later.] With diminished representation of their concerns on the executive team and the board, OpenAI's Safety clan was now significantly weakened. By April 2024, with the conclusion of the investigation, many, especially those [who were the most pessimistic about AI's catastrophic potential], were growing disillusioned and departing.

As the Safety clan's numbers depleted, the rest of the company was back to advancing its vision for *Her*. It now had all the ingredients: global brand recognition, real data on user behaviors from ChatGPT and its other products, and its newly trained model, Scallion [which was surpassing performance expectations.]

[Following The Blip, Altman's behavior became noticeably less measured and arrogant, often betraying his own carefully manicured persona of restraint by tweeting out boastful remarks to goad OpenAI's competitors.]

The more OpenAI faced uphill challenges, the more Altman seemed to overcompensate with public declarations of its extraordinary success. The pattern was becoming so consistent it was turning into a signal: If Altman was being brazen and boastful, most likely something wasn't going well.

Pressures were coming from every direction. After The Blip, the board's phrasing "not consistently candid in his communications" had, as some in OpenAI expected, triggered several investigations from regulators and law enforcement, including one from the US Securities and Exchange Commission into whether company investors had been misled, according to *The Wall Street Journal*. In the same month, *The New York Times* filed its copyright infringement lawsuit, which added to a snowballing pile of other lawsuits from artists, writers, and coders over OpenAI's reaping hundreds of millions, then billions, of dollars from models trained without credit, consent, or compensation on their work and that were now being used to automate away their jobs.

The board crisis had emboldened [Altman's] hidden detractors to emerge from the woodwork. More media stories were coming out with fresh sources willing to characterize Altman as having had a long history of dishonesty, power grabbing, and self-serving tactics. More were reaching out to Annie and publishing her perspective. Then there was the continued relentlessness of travel and the constricting reality of a new level of fame, which, among other things, no longer allowed Altman to stay anonymous in public. "It's a strangely isolating way to live," he said on a podcast. "I didn't think I would not be able to go out to dinner in my own city." **(389, 390, 396, 398)**

[Following the departure of Sutskever, and then Jan Leike, OpenAI's head of Superalignment, shortly thereafter, the company entered its second major reputational crisis.]

On the face of it, Sutskever's and Leike's departures seemed like a natural continuation in the broader trend at the company: the steady exodus of Safety people. After the two leaders' exits, those departures would accelerate. Many of the rest of the former Superalignment staff would proceed to exit with the dissolution of their team.

But in a sharp twist, the exodus would mark the start of a new and intensified conflagration in the Boomer-Doomer fight over OpenAI. The fight wasn't over. The Doomers were just bringing it outside of the company.

[Following his resignation, Jan Leike tweeted on May 17th:] "I have been disagreeing with OpenAI leadership about the company's core priorities for quite some time, until we finally reached a breaking point." … "OpenAI is shouldering an enormous responsibility on behalf of all of humanity," he continued. "But over the past years, safety culture and processes have taken a backseat to shiny products."

Within hours of Leike's tweets … another tweet was going viral. Kelsey Piper, a senior writer at *Vox* for the EA-inspired section Future Perfect, had posted a new story. "When you leave OpenAI, you get an unpleasant surprise," she wrote in her tweet sharing the scoop, "a departure deal where if you don't sign a lifelong nondisparagement commitment, you lose all of your vested equity."

Touching vested equity was a glaring red line in Silicon Valley. The value of an employee's shares could often far exceed their cash compensation, making or breaking their financial future. In OpenAI's case, the company was also building what [insiders viewed] as the most powerful and existentially dangerous technology in the world. It was paramount … for former employees to have the right to criticize and pressure the company in public in order to help hold it accountable. But the threat of having one's financial security disappear overnight would do well to muzzle anyone.

With Piper's story out, OpenAI's Slack lit up. In a channel called #i- have-a-question, a place for employees to ask about anything, someone posted a link to Piper's tweet. "Is this accurate?"

[As OpenAI's leadership scrambled to convince its employees that the nondisparagement clause was some kind of oversight, the company entered its third reputational crisis that, like "The Blip," would internally take on the moniker of The Omnicrisis.] **(400–402)**

On May 20, Scarlett Johansson released a blistering statement [in response to a blog post that OpenAI published a day prior to preemptively minimize speculation that one of GPT-4o's voice-mode personas, named Sky, was the voice of Johannson.]

All week, on top of everything else happening, OpenAI had been repeatedly fielding questions from journalists about the uncanny parallels between GPT-4o and the AI assistant Samantha, voiced by Johansson, in the movie *Her*. Behind the scenes, Johansson and her agent … had been pressing the company for the same clarifications. Publicly and privately, OpenAI had dismissed the similarities. When [the agent] called Altman demanding answers, Altman had been incredulous. Did they really think the voice sounded like her? Was she mad? he'd asked, according to an account from *The Wall Street Journal*.

[In a series of public statements, Johansson would detail how earlier in the year she had turned down several attempts by Altman to get the rights to use her voice before the company went public with GPT-4o at an AI expo, which had an avatar, Sky, whose voice so eerily matched the actress's that not even her friends could distinguish the difference, thereby raising the question if OpenAI trained their model of Johannson's vocal data without her consent.]

OpenAI hastily took down Sky and was back to defending itself. To its May 19 blog post, it added further elaboration and a statement from Altman. "The voice of Sky is not Scarlett Johansson's, and it was never intended to resemble hers," it said. "We are sorry to Ms. Johansson that we didn't communicate better."

After a string of other unflattering events, the Johansson scandal exploded publicly in a way that had not happened since the board crisis. It ripped through tech and policy corridors, igniting fresh speculation that Altman wasn't "consistently candid," in exactly the way the board had described him.

The Omnicrisis could have been a moment for OpenAI to engage in self-reflection. It was a prompt for the company to understand *why* exactly it had simultaneously lost the trust of employees, investors, and regulators as well as that of the broader public. Only then, maybe, just maybe, it would have begun to realize that both The Blip and the Omnicrisis were one and the same: the convulsions that arise from the deep systemic instability that occurs when an empire concentrates so much power, through so much dispossession, leaving the majority grappling with a loss of agency and material wealth and a tiny few to vie fiercely for control.

Instead, OpenAI chose to fortify itself against the criticism. Altman would repeat to employees, as he always had, that the Omnicrisis and The Blip were just the strange and expected moments of madness on the company's high-stakes noble quest to AGI. "As we all kind of feel that we're getting closer to finding our way to these powerful systems," he said during the May 15 all-hands, "the level of stress and tension will internally, externally, directed at us, emanating from us—that keeps going, that keeps increasing." The best way to manage it would be for OpenAI to double down on its PR, entrench its relationships with governments, and hold steadfast to its convictions in its vision. "I think on the whole we're quite good at that," he added, "but we will be tested again here." **(403–404, 410)**

Key words/phrases

412 (PDF) 399 (Book)	"Altman once remarked onstage that the best book he'd read the previous year in 2018 was The Mind of Napoleon, a more than three-hundred-page compilation of quotes from Napoleon Bonaparte, the French military leader who led a coup to seize control of the French government, installed himself as France's emperor, and subsequently sought to conquer Europe. "Obviously deeply flawed human, but man, impressive," Altman said. "What kinds of insights did he have?" asked Tyler Cowen, an economics professor at George Mason University, who was hosting the event. "His incredible understanding of human psychology," replied Altman, who was weeks away from switching to OpenAI full time and still president of YC. "That is something we see among many of our best founders."	
413 (PDF) 400 (Book)	"First, the mission centralizes talent by rallying them around a grand ambition, exactly in the way John McCarthy did with his coining of the phrase artificial intelligence. "The most successful founders do not set out to create companies," Altman reflected on his blog in 2013. "They are on a mission to create something closer to a religion, and at some point it turns out that forming a company is the easiest way to do so." Second, the mission centralizes capital and other resources while eliminating roadblocks, regulation, and dissent. Innovation, modernity, progress—what wouldn't we pay to achieve them? This is all the more true in the face of the scary, misaligned corporate and state competitors that supposedly exist. "Who will control the future of AI?" wrote Altman in a July 2024 op-ed for *The Washington Post* amid the aftershocks of the Omnicrisis. "Will it be one in which the United States and allied nations advance a global AI that spreads the technology's benefits and opens access to it, or an authoritarian one, in which nations or movements that don't share our values use AI to cement and expand their power?" Most consequentially, the mission remains so vague that it can be interpreted and reinterpreted—just as Napoleon did to the French Revolution's motto—to direct the centralization of talent, capital, and resources however the centralizer wants. What is beneficial? What is AGI? "I think it's a ridiculous and meaningless term," Altman told *The New York Times* just two days before the board fired him. "So I apologize that I keep using it." In this last ingredient, the creep of OpenAI has been nothing short of remarkable. In 2015, its mission meant being a nonprofit "unconstrained by a need to generate financial return" and open-sourcing research, as OpenAI wrote in its launch announcement. In 2016, it meant "everyone should benefit from the fruits of AI after its [sic] built, but it's totally OK to not share the science," as Sutskever wrote to Altman, Brockman, and Musk. In 2018 and 2019, it meant the creation of a capped profit structure "to marshal substantial resources" while avoiding "a competitive race without time for adequate safety precautions," as OpenAI wrote in its charter. In 2020, it meant walling off the model and building an "API as a strategy for openness and benefit sharing," as Altman wrote in response to my first profile. In 2022, it meant "iterative deployment" and racing as fast as possible to deploy ChatGPT. And in 2024, Altman wrote on his blog after the GPT-4o release: "A key part of our mission is to put very capable AI tools in the hands of people for free (or at a great price)."	Centralization is the key point that Hao wishes to surface in her critique of AI, highlighting the idea in relation to its economic, conceptual, and cultural senses. Curtailing other pathways for AI to develop forecloses other ways in which it could yield knowleges and possibilities
415 (PDF) 402 (Book)	"Behind the scenes, Altman was also laying the groundwork to entrench his own control. The board crisis had made clear that OpenAI's structure—a nonprofit governing a for-profit—had made his ouster as good as inevitable. The setup had not only enshrined the company's two countervailing forces —the Boomers and the Doomers—both vying for control of AI development but had also given the board the broad power to fire him based on their own, and not his, interpretation of whether or not he was best serving the mission. Such a conflict was bound to happen again if the structure stayed in place. On May 28, less than a week after executives sought to bring back Sutskever, an employee posted again in the Slack channel #i-have-a-question. "I don't know whether/how to ask this," the question began: Deep in the weeds of OpenAI's latest shareholder agreement were new details that seemed to allow for the dissolution of the nonprofit. "How solid is the non-profit?" the employee wrote. 'Is the plan to remain governed by a non-profit?'"	

418 (PDF) 405 (Book)

"At the same time, after over a year of work, OpenAI was still struggling to attain the desired performance for Orion to justify its release. The company was beginning to stare down the barrel of an uncomfortable prospect: Its tried-and-true formula of scaling no longer seemed to be enough to work; to advance its AI systems further, it likely needed fundamentally new research ideas. This was far easier said than done in general, but even more so after OpenAI had spent years orienting its hiring and team organization around exploiting existing research rather than exploring uncharted science."

421 (PDF) 409 (Book)

"It's hard to fully convey the tragedy of losing a language. For the same reasons AI researchers first gravitated toward language to build their technologies, the loss of a language extends far beyond the loss of a form of communication. Each language encodes within it rich histories, cultures, knowledge; it is the collective product of millions of people across time grasping for the sounds and written forms to capture the subtlest observations about the universe, about life, about the human experience; to share with one another stunning beauty and painful failure; to teach a child, to learn from an elder; to express love.

To lose a language is a global tragedy; it's also a personal one. To be severed from your inheritance and forced to preserve someone else's, or risk being beaten, is to establish, in one of the rawest ways possible, a clear hierarchy between whose history, whose culture, whose knowledge deserves to be passed down and whose is so insignificant it deserves to be erased."

Recalls the efforts to economically quantify the natural environment and how things that can't be quantified become trashed – in other words, we need an actuarial AI.

423 (PDF) 411 (Book)

"This meant that even before embarking on the project, they would get permission from the Māori community and their elders, asking them if the endeavor was even something they wanted; to collect the training data, they would seek contributions only from people who fully understood what the data would be used for and were willing to participate; to maximize the model's benefit, they would listen to the community for what kinds of language-learning resources would be most helpful; and once they had the resources, they would also buy their own on-site Nvidia GPUs and servers to train their models without a dependency on any tech giant's cloud.

Most crucially, Te Hiku would create a process by which the data it collected would continue to be a resource for future benefit but never be co-opted for projects that the community didn't consent to, that could exploit and harm them, or otherwise infringe on their rights. Based on the Māori principle of *kaitiakitanga*, or guardianship, the data would stay under Te Hiku's stewardship rather than be posted freely online; Te Hiku would then license it only to organizations that respected Māori values and intended to use it for projects that the community agreed to and found helpful."

Task oriented vs. God oriented tool.

426 (PDF) 414 (Book)

"After Timnit Gebru was ousted from Google, she founded a nonprofit in December 2021 to continue her research. She named it DAIR, the Distributed AI Research Institute—"distributed" to defy centralization. "That was the first word that came to my mind," Gebru says. She imagined building a team of researchers from around the world who would stay embedded in their communities to bring the rich experiences and perspectives of their local contexts to the institute's work, while also using that work to benefit those communities. "Tech is impacting the whole world out of Silicon Valley, but the whole world is not getting a chance to impact tech," she says.

Alex Hanna, a sociologist and one of the Google coauthors on the "Stochastic Parrots" paper, became the first to join Gebru as DAIR's director of research. Hanna's first order of business was to write a research philosophy to further elaborate the ethos of the organization's work. To do so, Gebru and Hanna hired DAIR's third person, Milagros Miceli, another sociologist and computer scientist who had been conducting research into the AI industry's exploitative labor practices. Together they wrote their philosophy: "Our research is intended to benefit communities which are typically not served by AI and create pathways to refuse, interrogate, and reshape AI systems together."

They created seven pillars for the philosophy's implementation, including centering and forging meaningful relationships with communities affected by but not yet typically represented in AI research, treating them as true partners in the pursuit of knowledge production, fairly compensating any forms of labor involved in the creation of research and technologies, questioning the systems underpinning AI development that marginalize those who've always been historically marginalized, and working with those communities to dream up alternatives that could bit by bit remold the world toward one they wanted to inhabit.

From there Miceli embarked on a new research project to put their philosophy into practice. She created the Data Workers' Inquiry and invited data workers from around the world to formulate their own research questions about the data-annotation industry and how to make it better. Regardless of where they lived, she paid them a standard researcher's salary in Germany, where she is based, to reflect the value of the work they did: twenty-five euros an hour."

Intelligence
Leadership
Global