# Chapter 6

## Ascension

Early in his career, Altman observed that new CEOs only succeeded if they "refounded" the company. He did this with conviction at YC when he inherited the presidency. He created new programs, including a new fund, to expand the accelerator's support for startups at different stages. He moved into hard technologies—those that required ambitious scientific innovation, including nuclear fusion, quantum computing, and self-driving cars. Already a prestigious name brand, YC's sphere of influence grew from a couple hundred companies to thousands a year, turning it into a center of gravity in Silicon Valley. " The thing that I'm most proud of is we really built an empire," Altman said after stepping down as president.

The end of his YC era marked the start of his new era at OpenAI. In March 2019, as he transitioned to OpenAI full time, he quickly brought with him the same aggressive mindset that he'd used at YC. He didn't want OpenAI to be among the world's leading AI organizations; he wanted it to be the only one. For years, Altman had taught other founders through YC and other forums to model the startup game as a winner-takes-all competition. If a startup had any hope of succeeding, he told them, they had to move swiftly and relentlessly to beat and then continue to beat back their rivals.

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.

At YC, Altman pushed his fellow partners to keep growing the number of companies it funded by 10x. " And we will, over time, figure out how to get another 10x and then another 10x after that," Altman later said of his strategy at an event. "Someday we will fund all the companies in the world." " Sam was the first person I ever heard say that, because of the work the original founders had done, and because of the brand that YC had created, we were in fact a de facto monopoly in this space," says Geoff Ralston, Altman's YC successor.

At OpenAI, Altman planned to use the same strategy. In a memo he sent to the company in late 2019 to articulate his long-term vision, he emphasized that OpenAI needed to "be number one" in four categories by the end of 2020: technical results, compute, money (to acquire more compute), and preparation, meaning the safety and security of the organization as well as its resilience to high-stress situations.

The most important of these was the first one, he said. If OpenAI wanted a chance at fulfilling its mission, it needed to build beneficial AGI first, or be such a leader that it could still shape AGI development. "Though we in theory could slow down capability work," he wrote, referring to advancing technical results, "given the rate of progress other people are making, we likely are required to move very quickly on technical progress if we want to have a lot of influence over AGI." This would only become increasingly true as more and more competitors caught on to OpenAI's strategy and moved into the space.

"We still need many more 10x leaps to get to AGI," he added later in the memo. "We should always work towards dramatic results, not incremental improvements."

Crucial to this success formula were several other considerations. It would be paramount for OpenAI to keep Microsoft happy to maintain the lead in compute. If OpenAI was successful, Microsoft had agreed that it would give far more than $1 billion. “We would like Microsoft to be our major partner all the way through,” Altman said. “They have the capability of delivering us, for the next 5 years at least, the most powerful supercomputers in the world.” This meant shifting away from the days of OpenAI’s freewheeling academic research environment and toward focused commercialization efforts to deliver Microsoft benefits. If OpenAI had other research projects it wanted to pursue, it would then have the resources. “To paraphrase that famous Disney quote,” Altman wrote, “we should make more money so that we can do more research, not do more research so that we can make more money.”

Additionally, the company needed to start pulling back on transparency. “The infohazard risk of talking about AGI will keep getting higher as we make more progress,” Altman argued. It was time to restrict research publications and model deployments, adopt a stricter confidentiality policy, and reveal progress on only narrow skills rather than more general AI advancements. Separately, everyone also needed to begin acting under the assumption that “every decision we make and every conversation we have ends up investigated and reported on the front page of The New York Times.”

That said, “it still seems very important that the world thinks we are winning at something,” he said. This would make “key influencers in the world” more “willing to go well out of their way to help us,” and make global policymakers “at the level of Presidents or their designees” come to OpenAI “for answers when they need to make big decisions.” To that end, “we should probably plan to release at least one very impressive demonstration of progress each year.”

Finally, the company needed to start acting with more seriousness and more unity. 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)."

---

Drama was in fact already brewing. Various little rifts that had bubbled up across the company were beginning to coalesce into big ones. Once quick to call each other friends, Brockman and the Amodei siblings were now butting heads on a growing list of issues. Among them, Dario Amodei's deprioritization of the *Dota 2* work had frustrated Brockman, who believed Amodei hadn't taken his contributions seriously. Where *Dota 2* was once the most compute-heavy project, Brockman also chafed against Amodei's centralization of compute for Nest's work on GPT-3. The Amodei siblings, meanwhile, found Brockman difficult to work with and were unwilling to let him join in on their language model development.

The tensions created a break among the leaders that slowly extended to the people who were loyal to each one in the company. During the *Dota 2* project, Brockman had forged a familial bond with some members of his team through the intense working hours, high stress, and a spur-of-the-moment retreat in Hawai'i, growing especially close to Jakub Pachocki and Szymon Sidor, two Polish scientists who were roommates and best friends. Amodei's AI safety teams, and the core members of the Nest team in particular, formed another contingent, bound together by their shared concern, in varying degrees, of rogue AI and existential or other extreme risks. They kept their work insulated from the rest of the company, creating

private Slack channels and documents not accessible even to other executives. It frustrated many more people beyond Brockman as they felt similarly sidelined by the dwindling of their compute resources, along with their visibility into the company's core research.

Amodei's AI safety contingent, meanwhile, was also growing disquieted with some of Altman's behaviors. 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.

"We're all pragmatic people," a person in the group says. "We're obviously raising money; we're going to do commercial stuff. It might look very reasonable if you're someone who makes loads of deals like Sam, to be like, 'All right, let's make a deal, let's trade a thing, we're going to trade the next thing.' And then if you are someone like me, you're like, 'We're trading a thing we don't fully understand.' It feels like it commits us to an uncomfortable place."

This was against the backdrop of a growing paranoia over different issues across the company. Within the AI safety contingent, it centered on what they saw as strengthening evidence that powerful misaligned AI systems could lead to disastrous outcomes. One bizarre experience in particular had left several of them somewhat nervous. In 2019, on a model trained after GPT-2 with roughly twice the number of parameters, a group of researchers had begun advancing the AI safety work that Amodei had wanted: testing reinforcement learning from human feedback as a way to guide the model toward generating cheerful and positive content and away from anything offensive.

But late one night, a researcher made an update that included a single typo in his code before leaving the RLHF process to run overnight. That typo was an important one: It was a minus sign flipped to a plus sign that made the RLHF process work in reverse, pushing GPT-2 to generate *more* offensive content instead of less. By the next morning, the typo had

wreaked its havoc, and GPT-2 was completing every single prompt with extremely lewd and sexually explicit language. It was hilarious—and also concerning. After identifying the error, the researcher pushed a fix to OpenAI's code base with a comment: Let's not make a utility minimizer.

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. At times this rankled employees who were not American. During lunches, they would question, Why did it have to be a US organization? remembers a former employee. Why not one from Europe? Why *not* one from China?

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. On Fridays, employees would kick back after a long week for music and wine nights, unwinding to the soothing sounds of a rotating cast of colleagues playing the office piano late into the night.

The shift in gravity unsettled some people, heightening their anxiety about random and unrelated incidents. Once, a journalist tailgated someone inside the gated parking lot to gain access to the building. Another time, an employee found an unaccounted-for USB stick, stirring consternation about whether it contained malware files, a common vector of attack, and was some kind of attempt at a cybersecurity breach. After it was examined on an air-gapped computer, one completely severed from the internet, the USB turned out to be nothing. At least twice, Amodei also used an air-gapped computer to write critical strategy documents, connecting the machine directly to a printer to circulate only physical copies. He was paranoid about state actors stealing OpenAI's secrets and building their own powerful AI models for malicious purposes.

“No one was prepared for this responsibility,” one employee remembers. “It kept people up at night.”

Altman himself was paranoid about people leaking information. He privately worried about Neuralink staff, with whom OpenAI continued to share an office, now with more unease after Musk’s departure. Altman worried, too, about Musk, who wielded an extensive security apparatus including personal drivers and bodyguards. Keenly aware of the capability difference, Altman at one point secretly commissioned an electronic countersurveillance audit in an attempt to scan the office for any bugs that Musk may have left to spy on OpenAI.

To employees, Altman used the specter of US adversaries advancing AI research faster than OpenAI to rationalize why the company needed to be less and less open while working as fast as possible. “We must hold ourselves responsible for a good outcome for the world,” he wrote in his vision document. “On the other hand, if an authoritarian government builds AGI before we do and misuses it, we will have also failed at our mission—we almost certainly have to make rapid technical progress in order to succeed at our mission.”

---

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? At a baseline, the executives agreed that they needed to lock down the model weights—the key information that could be used to replicate the fully trained versions of OpenAI’s deep learning models. If stolen, that would be bad because it could both empower bad actors and handicap OpenAI’s competitive advantage.

At first, without formal security staff, Altman deputized a member of the infrastructure team, which handled everything from the company’s GPUs to the office internet, to think about solutions for preventing model theft—not just from corporate or state-sponsored spies but also from OpenAI’s own employees. In cybersecurity, protecting against “insider threat” is relatively standard practice. Insiders could sabotage or steal OpenAI’s IP intentionally; they could also be tricked into giving it up. In

private, Altman acknowledged, after the point was raised, that someone like Sutskever could be vulnerable to the latter. The chief scientist was a logical target for bad actors: He was the archetype of a brainiac scientist who wasn't the most streetwise, and he ranked highly within the organization and had top access to information.

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.

Hidden from view of most employees, digital security increased with the installation of corporate-monitoring software. In the background, enhancements were also made to physical security. The gates to the office parking lot were fortified. Within the office, several doors with keypads were programmed to have "distress passwords," special codes that could be punched in to trigger a secret alarm that would alert relevant security personnel of an in-person threat. Quotes were sought from vendors about how much it would cost to reinforce a server room to withstand a machine gun, though that idea was subsequently dropped.

In the vision memo, Altman noted the divisions that were developing in the company from the heightening stress. "We have (at least) three clans at OpenAI—to caricature-ize them, let's say exploratory research, safety, and startup." The Exploratory Research clan was about advancing AI capabilities, the Safety clan about focusing on responsibility, and the Startup clan about moving fast and getting things done.

Per Altman, each of these clans had important values that the company needed to preserve: the "we will pursue important new ideas even if we fail

many times” of Exploratory Research; the “we will have an unwavering commitment to doing the right thing” of Safety; and the “we’ll figure out a way to make it happen” of Startup. “We have to continue to avoid tribal warfare,” he said. “To succeed, we need these three clans to unite as one tribe—while maintaining the strengths of each clan—working towards AGI that maximally benefits humanity.”

Though Altman never name-checked anyone, employees read between the lines. Sutskever was the face of Exploratory Research; Amodei and his AI safety contingent focused on extreme risks constituted Safety; Brockman was the champion of Startup. Soon after, the pandemic hit, and everyone began working remotely, making it far easier for the clans to isolate themselves from one another.

---

Amodei pushed his team to move quickly. As they had done with GPT-2, they trained iteratively larger models in the ascension to a full ten-thousand-GPU model with 175 billion parameters, naming them alphabetically after scientists: *ada* for the smallest model, referring to English mathematician Ada Lovelace, widely credited as the first computer programmer; *babbage* for English inventor Charles Babbage, who conceived the first digital computer for which Lovelace would propose her program; *curie* for Polish French physicist and chemist Marie Curie, the first woman to win the Nobel Prize and win it twice; and *davinci* for Leonardo. The exercise was both to continue validating whether scaling laws still held at fundamentally larger scale and, more practically, to work gradually through the hardware and data challenges at each new level. On a regular basis, the Nest team would give the company an update on its progress, to growing excitement. “It’s hard to overstate how insane that was to see,” remembers one researcher. “I’d never seen anything like that in my life.”

In parallel, Altman and Brockman developed a plan for commercialization. In late January 2020, Brockman began writing the first lines of code for an application programming interface, or API, for GPT-3. The API would give companies and developers access to the model’s capabilities without giving them access to the model weights and allow them to incorporate the technology into their own consumer-facing products. The company split into two divisions. Mira Murati was promoted

to VP of a new Applied division for overseeing the API and commercialization strategy. Under her, Peter Welinder, who had been leading the robotics team, was shifted to leading product; Fraser Kelton, who had cofounded an AI startup acquired by Airbnb, and Katie Mayer, who had worked at Leap Motion, were hired to respectively manage new product and engineering teams. Everyone not in Applied by association became the Research division.

That split deepened a fault line that Altman had identified. The formation of the Applied division brought in a small but growing group of people hailing from other startups that strengthened the Startup clan. While the Exploratory Research clan viewed this with some ambivalence about whether OpenAI would become just another Silicon Valley product company, it triggered increasingly impassioned opposition from Amodei and his Safety clan also sitting within the Research division.

To many in Safety, releasing GPT-3 in short order via an API, or any other means, undermined the lead time—the whole point of the accelerated scaling—that OpenAI would have to perfect the safety of the model. The Applied division, whose entire purpose was to find early solutions for making money from OpenAI's technologies, which in their view required releasing them in the near term, disagreed. The API as they saw it also gave OpenAI the most controlled mechanism of any release strategy, allowing the company to be selective about whom to give access to and collecting invaluable data points for understanding how the model could be used or abused by people. In all-hands meetings, Altman played both sides: The API would ultimately help each group achieve what they wanted; bringing in some revenue would allow OpenAI to invest even more in AI safety research.

As GPT-3 finished training, employees began playing with the model internally. They tested the bounds of its capabilities and tinkered with the first version of the API. The company held a hackathon where employees riffed on different application ideas. But with every new prototype, tensions worsened. Where the Applied division, and many in Exploratory Research, viewed the demonstrations with mounting excitement, many in Safety saw them as yet further evidence that releasing the model without comprehensive testing and additional research could risk devastating outcomes.

One capability proved particularly polarizing: GPT-3's code-generation abilities. It hadn't been part of the Nest team's intentions, but in scraping links on Reddit and using Common Crawl for training data, they had captured scattered lines of code from engineers posting their programs on various online forums to ask questions or share tips, leading the model to have an increased facility for programming languages. The development thrilled many in Exploratory Research, just as it did the Applied division. Not only was it an impressive technical milestone, it also had potential as a tool to accelerate the company's productivity in AI research and to make GPT-3 into a more compelling product. For the same reason, some in Safety panicked. If an AI system could use its own code-generation skills to tweak itself, it could accelerate the timeline to more powerful capabilities, increase the risk of it subverting human control, and amplify the chances of extremely harmful or existential AI risks.

Sutskever and Wojciech Zaremba, one of the founding members whom Musk had pressed during a meeting, would subsequently form a team to create a model designed specifically for code generation. But during a meeting to kick off the project, the two learned that Amodei already had his own plans for developing a code-generation model and didn't see a need to merge efforts. Despite his concerns, Amodei believed, as with GPT-3, that the best way to mitigate the possible harms of code generation was simply to build the model faster than anyone else, including even the other teams at OpenAI who he didn't believe would prioritize AI safety, and use the lead time to conduct research on de-risking the model. Much to the confusion of other employees, the two teams continued to work on duplicate code-generation efforts. "It just seemed from the outside watching this that it was some kind of crazy *Game of Thrones* stuff," a researcher says.

The deadlock around releasing GPT-3 via the API continued until late spring. Safety continued to push for paramount caution based on fears of accelerating extreme AI risks, arguing for the company to delay its release as long as possible. The Applied division continued with preparing for the API launch, arguing that the best way to improve the model was for it to have contact with the real world. Around the same time, new concerns emerged from a third group of employees worried about the impact that spectacular text-generation abilities could have in the midst of major political, social, and economic upheaval in the US. By May 2020, the

pandemic had already created a faster rise in unemployment than during the Great Recession. In the same month, Derek Chauvin, a police officer in Minneapolis, murdered George Floyd, a forty-six-year-old Black man, setting off massive Black Lives Matter protests around the country and the rest of the world. The team was also concerned about the impending US presidential election.

But rumors began to spread within OpenAI that Google could soon release its own large language model. The possibility was plausible. Google had published research at the start of the year about a new chatbot called Meena, built on a large language model with 1.7 times more parameters than GPT-2. The company could very well be working on scaling that model to roughly the size of GPT-3. The rumors sealed the deal for the API launch: If a model just as large would soon exist in the world, Safety felt less of a reason to hold back GPT-3.

In June, the company announced the API and set up an application form for people to request early access, prioritizing larger enterprises that the Applied division felt could be trusted to handle the technology responsibly. The company also maintained a big spreadsheet for employees to put down the names of anyone they wanted to jump the queue, including family, friends, and their favorite celebrities.

Google's rumored model never materialized. The tech giant had indeed begun working on a larger model than Meena, known as LaMDA, to produce a better chatbot—but it was still modestly smaller than GPT-3, and the company would ultimately decide not to release it until after ChatGPT. Google's executives determined that LaMDA didn't meet the company's ethical AI standards. Some employees also worried about repeating an infamous Microsoft scandal: In 2016, Microsoft had released an AI-powered chatbot known as Tay that quickly turned racist and misogynistic, and espoused support for Hitler, after users repeatedly prompted the chatbot to repeat inappropriate and offensive things. The GPT-3 API release wouldn't be the last decision that OpenAI would make to push out its technology based on an inflated fear of competition.

---

Just as ChatGPT would make OpenAI an instant household name, GPT-3 was that moment within AI and tech circles. In late 2022, ChatGPT would

add key improvements and features to the GPT-3 experience that would transform it into a globally viral product, including a consumer-friendly web interface, conversational abilities, more safety mechanisms, and a free version. But many of the core capabilities that the broader public would experience with the chatbot then, developers were already experiencing with the API in 2020, two years earlier. With the same awe and wonder, developers couldn't believe it.

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.

At NeurIPS that year, OpenAI's paper explaining its work on the model won one of the top research awards, surprising employees and establishing the lab's status as a leading organization. The effect was as the leadership team had predicted. OpenAI's new stature made it easier to recruit and retain talent, significantly helped along by the capital raised from OpenAI LP, which allowed the company to finally compete with Google and DeepMind on salaries.

In October 2020, with OpenAI's elevating recognition, Altman hired Steve Dowling, a seasoned executive who'd led communications at Apple, to be OpenAI's new VP of communications. He also placed Dowling in charge of government relations, emphasizing the importance of educating policymakers about AI and making them aware of the coming capabilities. After Jack Clark's departure, Dowling would bring on Anna Makanju, a highly respected former adviser in the Obama administration who had also worked on policy at Facebook and Musk's Starlink, to take over policy and global affairs.

Eager to ride GPT-3's momentum, the Applied division brainstormed ways to develop and expand its commercialization strategy. But seemingly at every turn, the Safety clan continued to put up resistance. 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. At one point, Welinder, who would become VP of product, commented bitterly that every conversation felt like a reenactment of a 1944 US intelligence manual about nonviolent sabotage. One section of the pamphlet, declassified in 2008, lists simple instructions for how to destabilize and undermine the productivity of an organization, including:

- Talk as frequently as possible and at great length.
- Bring up irrelevant issues as frequently as possible.
- Haggle over precise wordings of communications, minutes, resolutions.
- Refer back to matters decided upon at the last meeting and attempt to re-open the question of the advisability of that decision.
- Ask endless questions.

The animosity permeated outside meetings. To people in the Applied division, it felt like every digital communications channel was being co-opted into a battleground. A post from a product person in Slack could trigger dozens, if not more, concerned replies from people in Safety. A Google doc from Murati or Welinder sharing new thoughts on commercialization strategy could receive so many comments that the whole thing would appear covered in yellow highlights. The fact that GPT-3 was out in the world and the world hadn't ended made many in Applied also feel that the Safety clan was being hysterical for reasons that seemed completely detached from reality. To Safety, it was a matter of principle and precedent. OpenAI needed to establish rigorous norms and uphold itself to higher standards than might appear necessary in the moment. Once the stakes got

higher—and, Safety believed, they could get higher quickly and unpredictably—OpenAI's preparation would be the difference between its technologies bringing overwhelming harm or overwhelming benefit.

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. In late summer of 2020, the company dissolved its robotics team. Most of the robotics staff shifted to GPT projects; two mechanical engineers were laid off. By September, Microsoft announced that it would exclusively license GPT-3 from OpenAI, dramatically increasing the model's distribution. In addition to OpenAI continuing to offer GPT-3 through its API, Microsoft would now get full access to the model weights to embed and repurpose as it wished in its products and services, including to deliver in its own GPT-3 API on Azure.

As employees celebrated OpenAI's newfound popularity remotely from their homes, Dario and Daniela Amodei, who was now VP of safety and policy, Jack Clark, and several of the AI safety researchers who served as the core members of the Nest team suddenly fell quiet on Slack. 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."

As the group grappled with their disempowerment, they coalesced around a new idea. Dario Amodei first floated it to Jared Kaplan, a close friend from grad school and former roommate who worked part time at OpenAI and had led the discovery of scaling laws, and then to Daniela, Clark, and a small group of key researchers, engineers, and others loyal to

his views on AI safety. Did they really need to keep fighting for better AI safety practices at OpenAI? he asked. Could they break off to pursue their own vision? After several discussions, the group determined that if they planned to leave, they needed to do so imminently. With the way scaling laws were playing out, there was a narrowing window in which to build a competitor. “Scaling laws mean the requirements for training these frontier things are going to be going up and up and up,” says one person who parted with Amodei. “So if we wanted to leave and do something, we’re on a clock, you know?”

In late 2020, employees logged on to a video call for an all-hands meeting. Altman passed the mic to Dario Amodei, who was twirling and tugging his curly hair, as he often did, with a restless energy. He read a canned statement announcing that he, Daniela, and several others were leaving to form their own company. Altman then asked everyone quitting to leave the meeting. In May of the following year, the departed group announced a new public benefit corporation: Anthropic.

Anthropic people would later frame The Divorce, as some called it, as a disagreement over OpenAI’s approach to AI safety. While this was true, it was also about power. As much as Dario Amodei was motivated by a desire to do what was right within his principles and to distance himself from Altman, he also wanted greater control of AI development to pursue it based on his own values and ideology. He and the other Anthropic founders would build up their own mythology about why Anthropic, not OpenAI, was a better steward of what they saw as the most consequential technology. In Anthropic meetings, Amodei would regularly punctuate company updates with the phrase “unlike Sam” or “unlike OpenAI.” But in time, Anthropic would show little divergence from OpenAI’s approach, varying only in style but not in substance. Like OpenAI, it would relentlessly chase scale. Like OpenAI, it would breed a heightened culture of secrecy even as it endorsed democratic AI development. Like OpenAI, it would talk up cooperation when the very premise of its founding was rooted in rivalry.