Artificial intelligence / Machine learning

The messy, secretive reality behind OpenAI's bid to save the world

The AI moonshot was founded in the spirit of transparency. This is the inside story of how competitive pressure eroded that idealism.

by **Karen Hao** February 17, 2020

Greg Brockman, co-founder and CTO; Ilya Sutskever, co-founder and chief scientist; and Dario Amodei, research director.

CHRISTIE HEMM KLOK

Every year, OpenAI's employees vote on when they believe artificial general intelligence, or AGI, will finally arrive. It's mostly seen as a fun way to bond, and their estimates differ widely. But in a field that still debates whether human-like autonomous systems are even possible, half the lab bets it is likely to happen within 15 years.

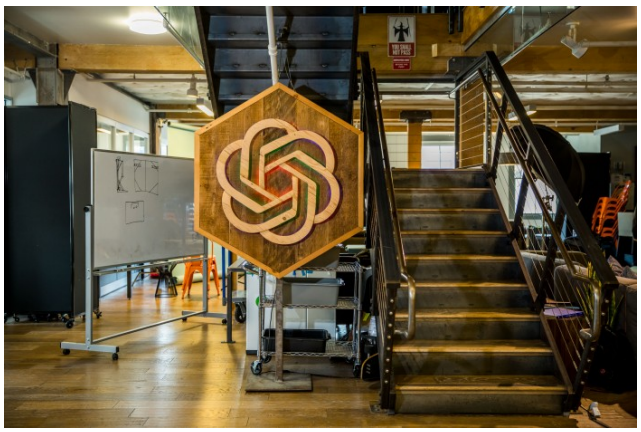
In the four short years of its existence, OpenAI has become one of the leading AI research labs in the world. It has made a name for itself producing consistently headline-grabbing research, alongside other AI heavyweights like Alphabet's DeepMind. It is also a darling in Silicon Valley, counting Elon Musk and legendary investor Sam Altman among its founders.

Above all, it is lionized for its mission. Its goal is to be the first to create AGI—a machine with the learning and reasoning powers of a human mind. The purpose is not world domination; rather, the lab wants to ensure that the technology is developed safely and its benefits distributed evenly to the world.

The implication is that AGI could easily run amok if the technology's development is left to follow the path of least resistance. Narrow intelligence, the kind of clumsy AI that surrounds us today, has already served as an example. We now know that algorithms are biased and fragile; they can perpetrate great abuse and great deception; and the expense of developing and running them tends to

OpenAI wants to be that shepherd, and it has carefully crafted its image to fit the bill. In a field dominated by wealthy corporations, it was founded as a nonprofit. Its first announcement said that this distinction would allow it to "build value for everyone rather than

shareholders." Its charter—a document so sacred that employees' pay is tied to how well they adhere to it—further declares that OpenAI's "primary fiduciary duty is to humanity." Attaining AGI safely is so important, it continues, that if another organization were close to getting there first, OpenAI would stop competing with it and collaborate instead. This alluring narrative plays well with investors and the media, and in July Microsoft injected the lab with a fresh $1 billion.



OpenAI's logo hanging in its office.
Christie Hemm Klok

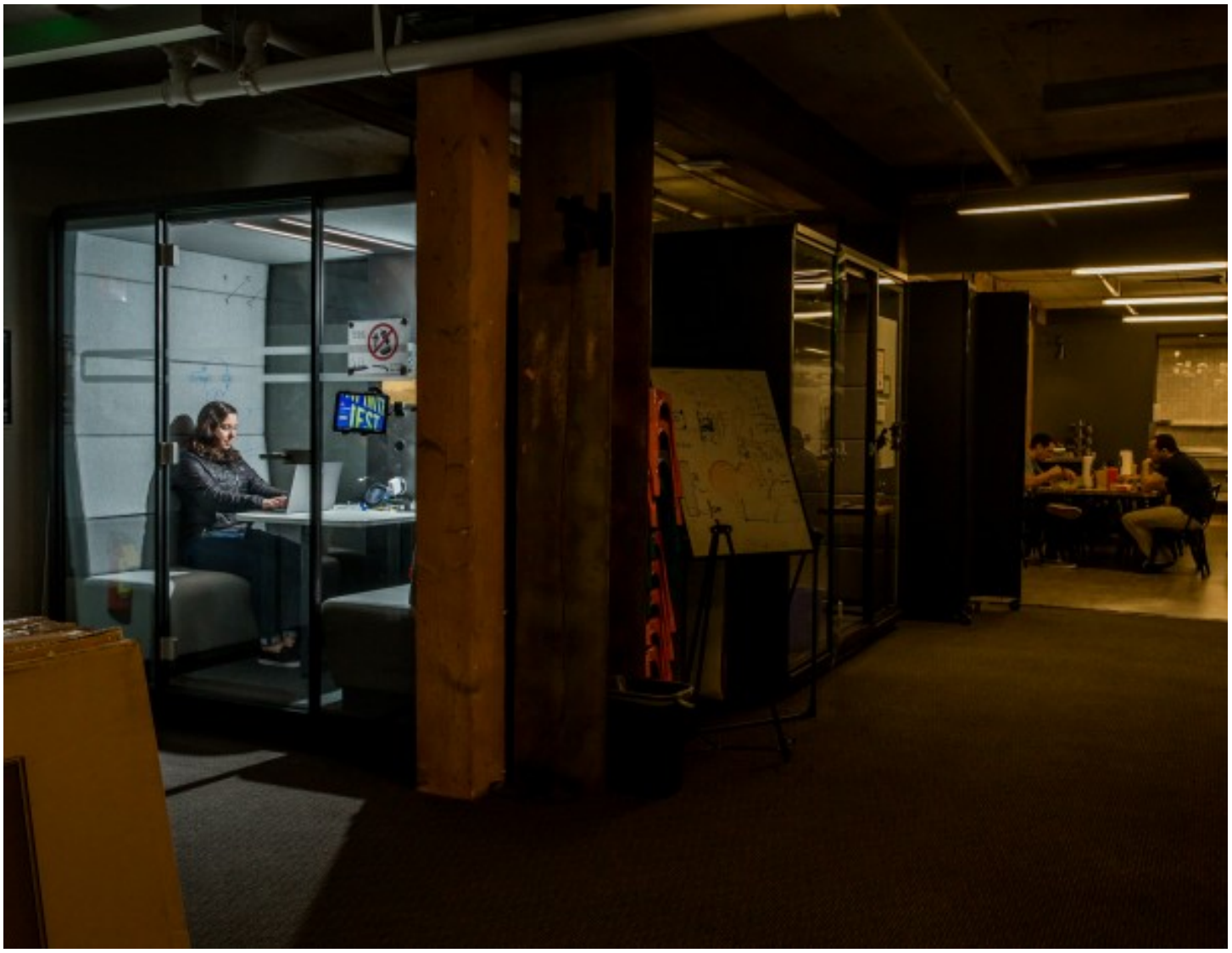
But three days at OpenAI's office—and nearly three dozen interviews with past and current employees, collaborators, friends, and other experts in the field—suggest a different picture. There is a misalignment between what the company publicly espouses and how it operates behind closed doors. Over time, it has allowed a fierce competitiveness and mounting pressure for ever more funding to erode its founding ideals of transparency, openness, and collaboration. Many who work or worked for the company insisted on anonymity because they were not authorized to speak or feared retaliation. Their accounts suggest that OpenAI, for all its noble aspirations, is obsessed with maintaining secrecy, protecting its image, and retaining the loyalty of its employees.

Since its earliest conception, AI as a field has strived to understand human-like intelligence and then re-create it. In 1950, Alan Turing, the renowned English mathematician and computer scientist, began a paper with the now-famous provocation "Can machines think?" Six years later, captivated by the nagging idea, a group of scientists gathered at Dartmouth College to formalize the discipline.

"It is one of the most fundamental questions of all intellectual history, right?" says Oren Etzioni, the CEO of the Allen Institute for Artificial Intelligence (AI2), a Seattle-based nonprofit AI research lab. "It's like, do we understand the origin of the universe? Do we understand matter?"

The trouble is, AGI has always remained vague. No one can really describe what it might look like or the minimum of what it should do. It's not obvious, for instance, that there is only one kind of general intelligence; human intelligence could just be a subset. There are also differing opinions about what purpose AGI could serve. In the more romanticized view, a machine intelligence unhindered by the need for sleep or the inefficiency of human communication could help solve complex challenges like climate change, poverty, and hunger.

But the resounding consensus within the field is that such advanced capabilities would take decades, even centuries—if indeed it's possible to develop them at all. Many also fear that pursuing this goal overzealously could backfire. In the 1970s and again in the late '80s and early '90s, the field overpromised and underdelivered. Overnight, funding dried up, leaving deep scars in an entire generation of researchers. "The field felt like a backwater," says Peter Eckersley, until recently director of research at the industry group Partnership on AI, of which OpenAI is a member.



:t floor named Infinite Jest.

Against this backdrop, OpenAI entered the world with a splash on December 11, 2015. It wasn't the first to openly declare it was pursuing AGI; DeepMind had done so five years earlier and had been acquired by Google in 2014. But OpenAI seemed different. For one thing, the sticker price was shocking: the venture would start with $1 billion from private investors, including Musk, Altman, and PayPal cofounder Peter Thiel.

The star-studded investor list stirred up a media frenzy, as did the impressive list of initial employees: Greg Brockman, who had run technology for the payments company Stripe, would be chief technology officer; Ilya Sutskever, who had studied under AI pioneer Geoffrey Hinton, would be research director; and seven researchers, freshly graduated from top universities or plucked from other companies, would compose the core technical team. (Last February, Musk announced that he was parting ways with the company over disagreements about its direction. A month later, Altman stepped down as president of startup accelerator Y Combinator to become OpenAI's CEO.)

But more than anything, OpenAI's nonprofit status made a statement. "It'll be important to have a leading research institution which can prioritize a good outcome for all over its own self-interest," the announcement said. "Researchers will be strongly encouraged to publish their work, whether as papers, blog posts, or code, and our patents (if any) will be shared with the world." Though it never made the criticism explicit, the implication was clear: other labs, like DeepMind, could not serve humanity because they were constrained by commercial interests. While they were closed, OpenAI would be *open*.

closely followed the lab's journey.

At the intersection of 18th and Folsom Streets in San Francisco, OpenAI's office looks like a mysterious warehouse. The historic building has drab gray paneling and tinted windows, with most of the shades pulled down. The letters "PIONEER BUILDING"—the remnants of its bygone owner, the Pioneer Truck Factory—wrap around the corner in faded red paint.

Inside, the space is light and airy. The first floor has a few common spaces and two conference rooms. One, a healthy size for larger meetings, is called A Space Odyssey; the other, more of a glorified phone booth, is called Infinite Jest. This is the space I'm restricted to during my visit. I'm forbidden to visit the second and third floors, which house everyone's desks, several robots, and pretty much everything interesting. When it's time for their interviews, people come down to me. An employee trains a watchful eye on me in between meetings.

On the beautiful blue-sky day that I arrive to meet Brockman, he looks nervous and guarded. "We've never given someone so much access before," he says with a tentative smile. He wears casual clothes and, like many at OpenAI, sports a shapeless haircut that seems to reflect an efficient, no-frills mentality.

The Pioneer Building.
wikimedia commons / tfinc

Brockman, 31, grew up on a hobby farm in North Dakota and had what he describes as a "focused, quiet childhood." He milked cows, gathered eggs, and fell in love with math while studying on his own. In 2008, he entered Harvard intending to double-major in math and computer science, but he quickly grew restless to enter the real world. He dropped out a year later, entered MIT instead, and then dropped out again within a matter of months. The second time, his decision was final. Once he moved to San Francisco, he never looked back.

Brockman takes me to lunch to remove me from the office during an all-company meeting. In the café across the street, he speaks about OpenAI with intensity, sincerity, and wonder, often drawing parallels between its mission and landmark achievements of science history. It's easy to appreciate his charisma as a leader. Recounting memorable passages from the books he's read, he zeroes in on the Valley's favorite narrative, America's race to the moon. ("One story I really love is the story of the janitor," he says, referencing a famous yet probably apocryphal tale. "Kennedy goes up to him and asks him, 'What are you doing?' and he says, 'Oh, I'm helping put a man on the moon!'") There's also the transcontinental railroad ("It was actually the last megaproject done entirely by hand … a project of immense scale that was totally risky") and Thomas Edison's incandescent lightbulb ("A committee of distinguished experts said 'It's never gonna work,' and one year later he shipped").

Greg Brockman, co-founder and CTO.
Christie Hemm Klok

Brockman is aware of the gamble OpenAI has taken on—and aware that it evokes cynicism and scrutiny. But with each reference, his message is clear: People can be skeptical all they want. It's the price of daring greatly.

Those who joined OpenAI in the early days remember the energy, excitement, and sense of purpose. The team was small—formed through a tight web of connections—and management stayed loose and informal. Everyone believed in a flat structure where ideas and debate would be welcome from anyone.

Musk played no small part in building a collective mythology. "The way he presented it to me was 'Look, I get it. AGI might be far away, but what if it's not?'" recalls Pieter Abbeel, a professor at UC Berkeley who worked there, along with several of his students, in the first two years. "'What if it's even just a 1% or 0.1% chance that it's happening in the next five to 10 years? Shouldn't we think about it very carefully?' That resonated with me," he says.

But the informality also led to some vagueness of direction. In May 2016, Altman and Brockman received a visit from Dario Amodei, then a Google researcher, who told them no one understood what they were doing. In an account published in the New Yorker, it wasn't clear the team itself knew either. "Our goal right now … is to do the best thing there is to do," Brockman said. "It's a little vague."

Nonetheless, Amodei joined the team a few months later. His sister, Daniela Amodei, had previously worked with Brockman, and he already knew many of OpenAI's members. After two years, at Brockman's request, Daniela joined too. "Imagine—we started with nothing," Brockman says. "We just had this ideal that we wanted AGI to go well."

Throughout our lunch, Brockman recites the charter like scripture, an explanation for every aspect of the company's existence.

within the field, they realized staying a nonprofit was financially untenable. The computational resources that others in the field were using to achieve breakthrough results were doubling every 3.4 months. It became clear that "in order to stay relevant," Brockman

says, they would need enough capital to match or exceed this exponential ramp-up. That required a new organizational model that could rapidly amass money—while somehow also staying true to the mission.

Unbeknownst to the public—and most employees—it was with this in mind that OpenAI released its charter in April of 2018. The document re-articulated the lab's core values but subtly shifted the language to reflect the new reality. Alongside its commitment to "avoid enabling uses of AI or AGI that harm humanity or unduly concentrate power," it also stressed the need for resources. "We anticipate needing to marshal substantial resources to fulfill our mission," it said, "but will always diligently act to minimize conflicts of interest among our employees and stakeholders that could compromise broad benefit."

"We spent a long time internally iterating with employees to get the whole company bought into a set of principles," Brockman says. "Things that had to stay invariant even if we changed our structure."

ɔdei, Jack Clark, Dario Amodei, Jeff Wu (technical staff member), Greg Brockman, Alec Radford (technical language team lead), ıff member), Ilya Sutskever, and Chris Berner (head of infrastructure).

That structure change happened in March 2019. OpenAI shed its purely nonprofit status by setting up a "capped profit" arm—a for-profit with a 100-fold limit on investors' returns, albeit overseen by a board that's part of a nonprofit entity. Shortly after, it announced Microsoft's billion-dollar investment (though it didn't reveal that this was split between cash and credits to Azure, Microsoft's cloud computing platform).

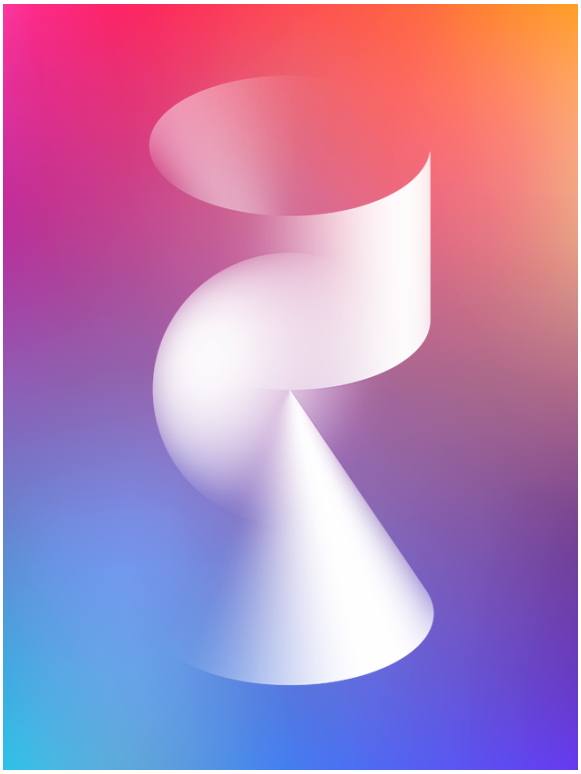
Predictably, the move set off a wave of accusations that OpenAI was going back on its mission. In a post on Hacker News soon after the announcement, a user asked how a 100-fold limit would be limiting at all: "Early investors in Google have received a roughly 20x return on their capital," they wrote. "Your bet is that you'll have a corporate structure which returns orders of magnitude more than Google ... but you don't want to 'unduly concentrate power'? How will this work? What exactly is power, if not the concentration of resources?"

paragraph of explanation.

The charter is the backbone of OpenAI. It serves as the springboard for all the lab's strategies and actions. Throughout our lunch, Brockman recites it like scripture, an explanation for every aspect of the company's existence. ("By the way," he clarifies halfway through one recitation, "I guess I know all these lines because I spent a lot of time really poring over them to get them exactly right. It's not like I was reading this before the meeting.")

How will you ensure that humans continue to live meaningful lives as you develop more advanced capabilities? "As we wrote, we think its impact should be to give everyone economic freedom, to let them find new opportunities that aren't imaginable today." *How will you structure yourself to evenly distribute AGI?* "I think a utility is the best analogy for the vision that we have. But again, it's all subject to the charter." *How do you compete to reach AGI first without compromising safety?* "I think there is absolutely this important balancing act, and our best shot at that is what's in the charter."

For Brockman, rigid adherence to the document is what makes OpenAI's structure work. Internal alignment is treated as paramount: all full-time employees are required to work out of the same office, with few exceptions. For the policy team, especially Jack Clark, the director, this means a life divided between San Francisco and Washington, DC. Clark doesn't mind—in fact, he agrees with the mentality. It's the in-between moments, like lunchtime with colleagues, he says, that help keep everyone on the same page.



APRIL 9, 2018 5 MINUTE READ
OpenAI

In many ways, this approach is clearly working: the company has an impressively uniform culture. The employees work long hours and talk incessantly about their jobs through meals and social hours; many go to the same parties and subscribe to the rational philosophy of "effective altruism." They crack jokes using machine-learning terminology to describe their lives: "What is your life a function of?" "What are you optimizing for?" "Everything is basically a minmax function." To be fair, other AI researchers also love doing this, but people familiar with OpenAI agree: more than others in the field, its employees treat AI research not as a job but as an identity. (In November, Brockman married his girlfriend of one year, Anna, in the office against a backdrop of flowers arranged in an OpenAI logo. Sutskever acted as the officiant; a robot hand was the ring bearer.)

But at some point in the middle of last year, the charter became more than just lunchtime conversation fodder. Soon after switching to a capped-profit, the leadership instituted a new pay structure based in part on each employee's absorption of the mission. Alongside columns like "engineering expertise" and "research direction" in a spreadsheet tab titled "Unified Technical Ladder," the last column outlines the culture-related expectations for every level. Level 3: "You understand and internalize the OpenAI charter." Level 5: "You ensure all projects you and your team-mates work on are consistent with the charter." Level 7: "You are responsible for upholding and improving the charter, and holding others in the organization accountable for doing the same."

The first time most people ever heard of OpenAI was on February 14, 2019. That day, the lab announced impressive new research: a model that could generate convincing essays and articles at the push of a button. Feed it a sentence from *The Lord of the Rings* or the start of a (fake) news story about Miley Cyrus shoplifting, and it would spit out paragraph after paragraph of text in the same vein.

But there was also a catch: the model, called GPT-2, was too dangerous to release, the researchers said. If such powerful technology fell into the wrong hands, it could easily be weaponized to produce disinformation at immense scale.

The backlash among scientists was immediate. OpenAI was pulling a publicity stunt, some said. GPT-2 was not nearly advanced enough to be a threat. And if it was, why announce its existence and then preclude public scrutiny? "It seemed like OpenAI was trying to capitalize off of panic around AI," says Britt Paris, an assistant professor at Rutgers University who studies AI-generated disinformation.

Jack Clark, policy director.
Christie Hemm Klok

By May, OpenAI had revised its stance and announced plans for a "staged release." Over the following months, it successively dribbled out more and more powerful versions of GPT-2. In the interim, it also engaged with several research organizations to scrutinize the algorithm's potential for abuse and develop countermeasures. Finally, it released the full code in November, having found, it said, "no strong evidence of misuse so far."

Amid continued accusations of publicity-seeking, OpenAI insisted that GPT-2 hadn't been a stunt. It was, rather, a carefully thought-out experiment, agreed on after a series of internal discussions and debates. The consensus was that even if it had been slight overkill this time, the action would set a precedent for handling more dangerous research. Besides, the charter had predicted that "safety and security concerns" would gradually oblige the lab to "reduce our traditional publishing in the future."

This was also the argument that the policy team carefully laid out in its six-month follow-up blog post, which they discussed as I sat in on a meeting. "I think that is definitely part of the success-story framing," said Miles Brundage, a policy research scientist, highlighting something in a Google doc. "The lead of this section should be: We did an ambitious thing, now some people are replicating it, and here are some reasons why it was beneficial."

But OpenAI's media campaign with GPT-2 also followed a well-established pattern that has made the broader AI community leery. Over the years, the lab's big, splashy research announcements have been repeatedly accused of fueling the AI hype cycle. More than once, critics have also accused the lab of talking up its results to the point of mischaracterization. For these reasons, many in the field have tended to keep OpenAI at arm's length.

search releases hang on its office wall.

This hasn't stopped the lab from continuing to pour resources into its public image. As well as research papers, it publishes its results in highly produced company blog posts for which it does everything in-house, from writing to multimedia production to design of the cover images for each release. At one point, it also began developing a documentary on one of its projects to rival a 90-minute movie about DeepMind's AlphaGo. It eventually spun the effort out into an independent production, which Brockman and his wife, Anna, are now partially financing. (I also agreed to appear in the documentary to provide technical explanation and context to OpenAI's achievement. I was not compensated for this.)

And as the blowback has increased, so have internal discussions to address it. Employees have grown frustrated at the constant outside criticism, and the leadership worries it will undermine the lab's influence and ability to hire the best talent. An internal document highlights this problem and an outreach strategy for tackling it: "In order to have government-level policy influence, we need to be viewed as the most trusted source on ML [machine learning] research and AGI," says a line under the "Policy" section. "Widespread support and backing from the research community is not only necessary to gain such a reputation, but will amplify our message." Another, under "Strategy," reads, "Explicitly treat the ML community as a comms stakeholder. Change our tone and external messaging such that we only antagonize them when we intentionally choose to."

There was another reason GPT-2 had triggered such an acute backlash. People felt that OpenAI was once again walking back its

But little did people know this wasn't the only time OpenAI had chosen to hide its research. In fact, it had kept another effort entirely secret.

There are two prevailing technical theories about what it will take to reach AGI. In one, all the necessary techniques already exist; it's just a matter of figuring out how to scale and assemble them. In the other, there needs to be an entirely new paradigm; deep learning, the current dominant technique in AI, won't be enough.

Most researchers fall somewhere between these extremes, but OpenAI has consistently sat almost exclusively on the scale-and-assemble end of the spectrum. Most of its breakthroughs have been the product of sinking dramatically greater computational resources into technical innovations developed in other labs.

Brockman and Sutskever deny that this is their sole strategy, but the lab's tightly guarded research suggests otherwise. A team called "Foresight" runs experiments to test how far they can push AI capabilities forward by training existing algorithms with increasingly large amounts of data and computing power. For the leadership, the results of these experiments have confirmed its instincts that the lab's all-in, compute-driven strategy is the best approach.

For roughly six months, these results were hidden from the public because OpenAI sees this knowledge as its primary competitive advantage. Employees and interns were explicitly instructed not to reveal them, and those who left signed nondisclosure agreements. It was only in January that the team, without the usual fanfare, quietly posted a paper on one of the primary open-source databases for AI research. People who experienced the intense secrecy around the effort didn't know what to make of this change. Notably, another paper with similar results from different researchers had been posted a few months earlier.

Ilya Sutskever, co-founder and chief scientist.
Christie Hemm Klok

Christie Hemm Klok

In the beginning, this level of secrecy was never the intention, but it has since become habitual. Over time, the leadership has moved away from its original belief that openness is the best way to build beneficial AGI. Now the importance of keeping quiet is impressed on those who work with or at the lab. This includes never speaking to reporters without the express permission of the communications team. After my initial visits to the office, as I began contacting different employees, I received an email from the head of communications reminding me that all interview requests had to go through her. When I declined, saying that this would undermine the validity of what people told me, she instructed employees to keep her informed of my outreach. A Slack message from Clark, a former journalist, later commended people for keeping a tight lid as a reporter was "sniffing around."

In a statement responding to this heightened secrecy, an OpenAI spokesperson referred back to a section of its charter. "We expect that safety and security concerns will reduce our traditional publishing in the future," the section states, "while increasing the importance of sharing safety, policy, and standards research." The spokesperson also added: "Additionally, each of our releases is run through an infohazard process to evaluate these trade-offs and we want to release our results slowly to understand potential risks and

years of research: an AI system trained on images, text, and other data using massive computational resources. A small team has been assigned to the initial effort, with an expectation that other teams, along with their work, will eventually fold in. On the day it was

announced at an all-company meeting, interns weren't allowed to attend. People familiar with the plan offer an explanation: the leadership thinks this is the most promising way to reach AGI.

The man driving OpenAI's strategy is Dario Amodei, the ex-Googler who now serves as research director. When I meet him, he strikes me as a more anxious version of Brockman. He has a similar sincerity and sensitivity, but an air of unsettled nervous energy. He looks distant when he talks, his brows furrowed, a hand absentmindedly tugging his curls.

Amodei divides the lab's strategy into two parts. The first part, which dictates how it plans to reach advanced AI capabilities, he likens to an investor's "portfolio of bets." Different teams at OpenAI are playing out different bets. The language team, for example, has its money on a theory postulating that AI can develop a significant understanding of the world through mere language learning. The robotics team, in contrast, is advancing an opposing theory that intelligence requires a physical embodiment to develop.

As in an investor's portfolio, not every bet has an equal weight. But for the purposes of scientific rigor, all should be tested before being discarded. Amodei points to GPT-2, with its remarkably realistic auto-generated texts, as an instance of why it's important to keep an open mind. "Pure language is a direction that the field and even some of us were somewhat skeptical of," he says. "But now it's like, 'Wow, this is really promising.'"

Over time, as different bets rise above others, they will attract more intense efforts. Then they will cross-pollinate and combine. The goal is to have fewer and fewer teams that ultimately collapse into a single technical direction for AGI. This is the exact process that OpenAI's latest top-secret project has supposedly already begun.

Dario Amodei, research director.
Christie Hemm Klok

The second part of the strategy, Amodei explains, focuses on how to make such ever-advancing AI systems safe. This includes making sure that they reflect human values, can explain the logic behind their decisions, and can learn without harming people in the process. Teams dedicated to each of these safety goals seek to develop methods that can be applied across projects as they mature. Techniques developed by the explainability team, for example, may be used to expose the logic behind GPT-2's sentence

"At some point we're going to build AGI, and by that time I want to feel good about these systems operating in the world," he says. "Anything where I don't currently feel good, I create and recruit a team to focus on that thing."

For all the publicity-chasing and secrecy, Amodei looks sincere when he says this. The possibility of failure seems to disturb him.

"We're in the awkward position of: we don't know what AGI looks like," he says. "We don't know when it's going to happen." Then, with careful self-awareness, he adds: "The mind of any given person is limited. The best thing I've found is hiring other safety researchers who often have visions which are different than the natural thing I might've thought of. I want that kind of variation and diversity because that's the only way that you catch everything."

The thing is, OpenAI actually has little "variation and diversity"—a fact hammered home on my third day at the office. During the one lunch I was granted to mingle with employees, I sat down at the most visibly diverse table by a large margin. Less than a minute later, I realized that the people eating there were not, in fact, OpenAI employees. Neuralink, Musk's startup working on computer-brain interfaces, shares the same building and dining room.

According to a lab spokesperson, out of the over 120 employees, 25% are female or nonbinary. There are also two women on the executive team and the leadership team is 30% women, she said, though she didn't specify who was counted among these teams. (All four C-suite executives, including Brockman and Altman, are white men. Out of over 112 employees I identified on LinkedIn and other sources, the overwhelming number were white or Asian.)

Daniela Amodei, head of people operations.
Christie Hemm Klok

In fairness, this lack of diversity is typical in AI. Last year a report from the New York–based research institute AI Now found that women accounted for only 18% of authors at leading AI conferences, 20% of AI professorships, and 15% and 10% of research staff at Facebook and Google, respectively. "There is definitely still a lot of work to be done across academia and industry," OpenAI's spokesperson said. "Diversity and inclusion is something we take seriously and are continually working to improve by working with initiatives like WiML, Girl Geek, and our Scholars program."

Indeed, OpenAI has tried to broaden its talent pool. It began its remote Scholars program for underrepresented minorities in 2018. But only two of the first eight scholars became full-time employees, even though they reported positive experiences. The most common reason for declining to stay: the requirement to live in San Francisco. For Nadja Rhodes, a former scholar who is now the lead machine-learning engineer at a New York–based company, the city just had too little diversity.

But if diversity is a problem for the AI industry in general, it's something more existential for a company whose mission is to spread the technology evenly to everyone. The fact is that it lacks representation from the groups most at risk of being left out.

Nor is it at all clear just *how* OpenAI plans to "distribute the benefits" of AGI to "all of humanity," as Brockman frequently says in citing its mission. The leadership speaks of this in vague terms and has done little to flesh out the specifics. (In January, the Future of Humanity Institute at Oxford University released a report in collaboration with the lab proposing to distribute benefits by distributing a percentage of profits. But the authors cited "significant unresolved issues regarding … the way in which it would be implemented.") "This is my biggest problem with OpenAI," says a former employee, who spoke on condition of anonymity.

Christie Hemm Klok

"They are using sophisticated technical practices to try to answer social problems with AI," echoes Britt Paris of Rutgers. "It seems like they don't really have the capabilities to actually understand the social. They just understand that that's a sort of a lucrative place to be positioning themselves right now."

Brockman agrees that both technical and social expertise will ultimately be necessary for OpenAI to achieve its mission. But he disagrees that the social issues need to be solved from the very beginning. "How exactly do you bake ethics in, or these other perspectives in? And when do you bring them in, and how? One strategy you could pursue is to, from the very beginning, try to bake in everything you might possibly need," he says. "I don't *think* that that strategy is likely to succeed."

The first thing to figure out, he says, is what AGI will even look like. Only then will it be time to "make sure that we are understanding the ramifications."

Last summer, in the weeks after the switch to a capped-profit model and the $1 billion injection from Microsoft, the leadership assured employees that these updates wouldn't functionally change OpenAI's approach to research. Microsoft was well aligned with the lab's values, and any commercialization efforts would be far away; the pursuit of fundamental questions would still remain at the core of the work.

For a while, these assurances seemed to hold true, and projects continued as they were. Many employees didn't even know what promises, if any, had been made to Microsoft.

But in recent months, the pressure of commercialization has intensified, and the need to produce money-making research no longer feels like something in the distant future. In sharing his 2020 vision for the lab privately with employees, Altman's message is clear: OpenAI needs to make money in order to do research—not the other way around.

This is a hard but necessary trade-off, the leadership has said—one it had to make for lack of wealthy philanthropic donors. By contrast, Seattle-based AI2, a nonprofit that ambitiously advances fundamental AI research, receives its funds from a self-sustaining (at least for the foreseeable future) pool of money left behind by the late Paul Allen, a billionaire best known for cofounding Microsoft.

But the truth is that OpenAI faces this trade-off not only because it's not rich, but also because it made the strategic choice to try to reach AGI before anyone else. That pressure forces it to make decisions that seem to land farther and farther away from its original intention. It leans into hype in its rush to attract funding and talent, guards its research in the hopes of keeping the upper hand, and chases a computationally heavy strategy—not because it's seen as the only way to AGI, but because it seems like the fastest.

Yet OpenAI is still a bastion of talent and cutting-edge research, filled with people who are sincerely striving to work for the benefit of humanity. In other words, it still has the most important elements, and there's still time for it to change.

Near the end of my interview with Rhodes, the former remote scholar, I ask her the one thing about OpenAI that I shouldn't omit from this profile. "I guess in my opinion, there's problems," she begins hesitantly. "Some of them come from maybe the environment it faces; some of them come from the type of people that it tends to attract and other people that it leaves out."

Update: *We made some changes to this story after OpenAI asked us to clarify that when Greg Brockman said he didn't think it was possible to "bake ethics in… from the very beginning" when developing AI, he intended it to mean that ethical questions couldn't be solved*

from the beginning, not that they couldn't be addressed from the beginning. Also, that after dropping out of Harvard he transferred straight to MIT rather than waiting a year. Also, that he was raised not "on a farm," but "on a hobby farm." Brockman considers this distinction important.

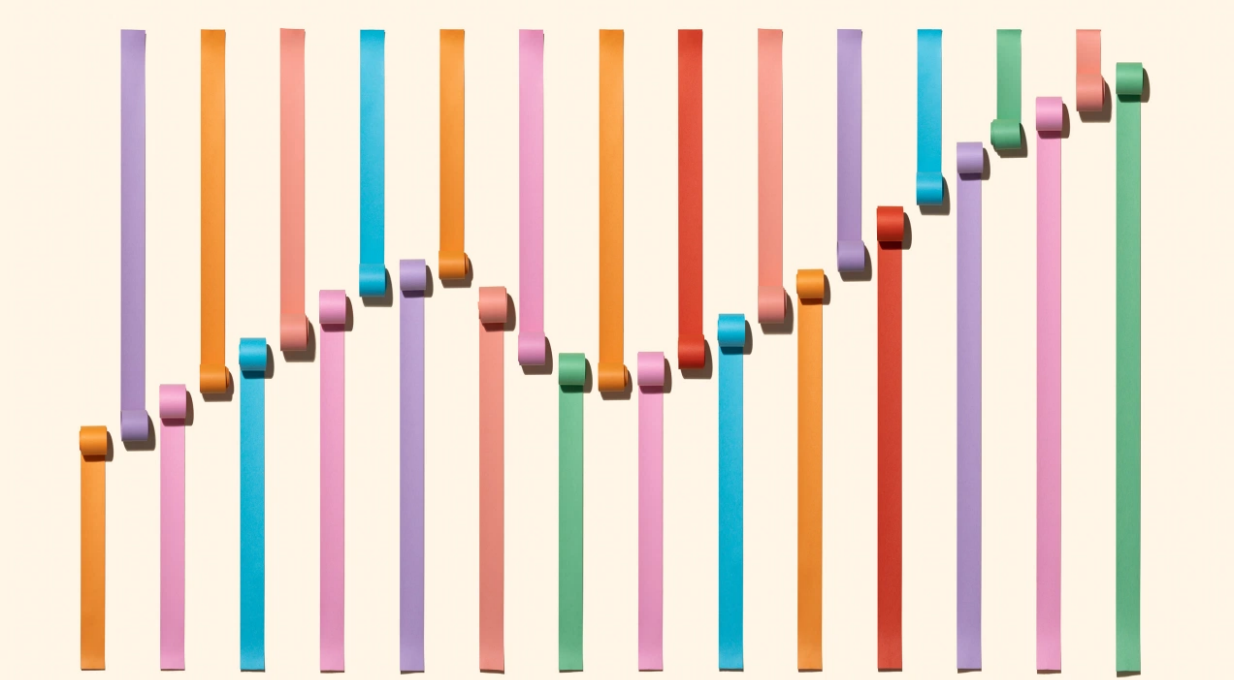
In addition, we have clarified that while OpenAI did indeed "shed its nonprofit status," a board that is part of a nonprofit entity still oversees it, and that OpenAI publishes its research in the form of company blog posts as well as, not in lieu of, research papers. We've also corrected the date of publication of a paper by outside researchers and the affiliation of Peter Eckersley (former, not current, research director of Partnership on AI, which he recently left).

Author Karen Hao

Computing 4 days

All together now: the most trustworthy covid-19 model is an ensemble

Combining a multitude of predictions and projections, modeling teams hone the uncertainty.



Computing 5 days

Chinese hackers posing as the UN Human Rights Council are attacking Uyghurs

Chinese-speaking hackers are targeting Uyghur Muslims with fake United Nations reports and phony support organizations, according to a new report.

Climate change 6 days

A startup using minerals to draw down CO2 has scored funding—and its first buyer

Heirloom Carbon Technologies believes it can pull off carbon removal for $50 a ton, and aims to remove one billion tons by 2035.

Humans and technology May 28

India is grappling with covid grief

India's wealth and tech divide mean that only some people get to grieve online in the country's deadly second covid wave.

01.
The world is waking up to India's plight—too late
Apr 27

02.
What India needs to get through its covid crisis
May 01

03.
How Indians are crowdsourcing aid as covid surges
Apr 28

MIT Technology Review

Space 6 days

Startup Phantom Space wants to be the

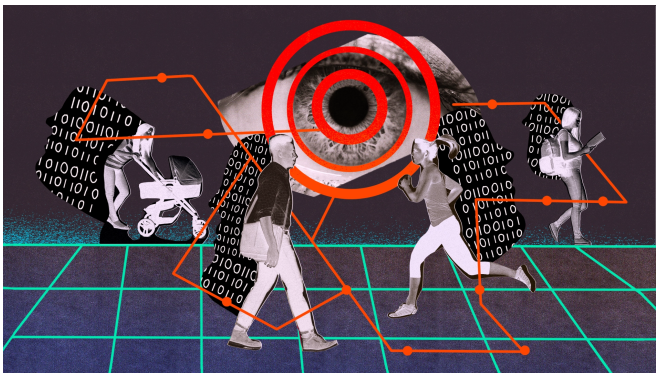
Henry Ford of rockets

The Arizona-based launch company wants to make enough rockets to launch 100 missions a year.

Opinion May 25

Collective data rights can stop big tech from obliterating privacy

Protecting individual data is not enough when the harms are collective



D-Lab project leads to solar career in Africa

Knocking on the door of innovation in Chile

View more