Ilya Sutskever, Sam Altman, Mira Murati, and Greg Brockman, of OpenAI PHOTOGRAPH: JESSICA CHOU

STEVEN LEVY BACKCHANNEL SEP 5, 2023 6:00 AM

What OpenAI Really Wants

The young company sent shock waves around the world when it released ChatGPT. But that was just the start. The ultimate goal: Change everything. Yes. *Everything*.

SAVE

0:00 / 1:06:14

Audio: Listen to this article.

THE AIR CRACKLES with an almost Beatlemaniac energy as the star and his entourage tumble into a waiting Mercedes van. They've just ducked out of one event and are headed to another, then another, where a frenzied mob awaits. As they careen through the streets of London—the short hop from Holborn to Bloomsbury—it's as if they're surfing one of civilization's before-and-after moments. The history-making force personified inside this car has captured the attention of the world. Everyone wants a piece of it, from the students who've waited in line to the prime minister.

Inside the luxury van, wolfing down a salad, is the neatly coiffed 38-year-old entrepreneur Sam Altman, cofounder of OpenAI; a PR person; a security specialist; and me. Altman is unhappily sporting a blue suit with a tieless pink dress shirt as he whirlwinds through London as part of a monthlong global jaunt through 25 cities on six continents. As he gobbles his greens—no time for a sit-down lunch today—he reflects on his meeting the previous night with French president Emmanuel Macron. Pretty good guy! And *very* interested in artificial intelligence.

As was the prime minister of Poland. And the prime minister of Spain.

Riding with Altman, I can almost hear the ringing, ambiguous chord that opens "A Hard Day's Night"—introducing the future. Last November, when OpenAI let loose its monster hit, ChatGPT, it triggered a tech explosion not seen since the internet burst into our lives. Suddenly the Turing test was history, search engines were endangered species, and no college essay could ever be trusted. No job was safe. No scientific problem was immutable.

Altman didn't do the research, train the neural net, or code the interface of ChatGPT and its more precocious sibling, GPT-4. But as CEO—and a dreamer/doer type who's

like a younger version of his cofounder Elon Musk, without the baggage—one news article after another has used his photo as the visual symbol of humanity's new challenge. At least those that haven't led with an eye-popping image generated by OpenAI's visual AI product, Dall-E. He is the oracle of the moment, the figure that people want to consult first on how AI might usher in a golden age, or consign humans to irrelevance, or worse.

Altman's van whisks him to four appearances that sunny day in May. The first is stealthy, an off-the-record session with the Round Table, a group of government, academia, and industry types. Organized at the last minute, it's on the second floor of a pub called the Somers Town Coffee House. Under a glowering portrait of brewmaster Charles Wells (1842–1914), Altman fields the same questions he gets from almost every audience. Will AI kill us? Can it be regulated? What about China? He answers every one in detail, while stealing glances at his phone. After that, he does a fireside chat at the posh Londoner Hotel in front of 600 members of the Oxford Guild. From there it's on to a basement conference room where he answers more technical questions from about 100 entrepreneurs and engineers. Now he's almost late to a mid-afternoon onstage talk at University College London. He and his group pull up at a loading zone and are ushered through a series of winding corridors, like the Steadicam shot in *Goodfellas*. As we walk, the moderator hurriedly tells Altman what he'll ask. When Altman pops on stage, the auditorium—packed with rapturous academics, geeks, and journalists—erupts.

Altman is not a natural publicity seeker. I once spoke to him right after *The New Yorker* ran a long profile of him. "Too much about me," he said. But at University College, after the formal program, he wades into the scrum of people who have surged to the foot of the stage. His aides try to maneuver themselves between Altman and the throng, but he shrugs them off. He takes one question after another, each time intently staring at the face of the interlocutor as if he's hearing the query for the first time. Everyone wants a selfie. After 20 minutes, he finally allows his team to pull him out. Then he's off to meet with UK prime minister Rishi Sunak.

Maybe one day, when robots write our history, they will cite Altman's world tour as a milestone in the year when everyone, all at once, started to make their own personal reckoning with the singularity. Or then again, maybe whoever writes the history of this moment will see it as a time when a quietly compelling CEO with a paradigm-

busting technology made an attempt to inject a very peculiar worldview into the global mindstream—from an unmarked four-story headquarters in San Francisco's Mission District to the entire world.

This article appears in the October 2023 issue. Subscribe to WIRED. PHOTOGRAPH: JESSICA CHOU

For Altman and his company, ChatGPT and GPT-4 are merely stepping stones along the way to achieving a simple and seismic mission, one these technologists may as well have branded on their flesh. That mission is to build artificial general intelligence—a concept that's so far been grounded more in science fiction than science—and to make it safe for humanity. The people who work at OpenAI are fanatical in their pursuit of that goal. (Though, as any number of conversations in the office café will confirm, the "build AGI" bit of the mission seems to offer up more raw excitement to its researchers than the "make it safe" bit.) These are people who do not shy from casually using the term "super-intelligence." They *assume* that AI's trajectory will surpass whatever peak biology can attain. The company's financial documents even stipulate a kind of exit contingency for when AI wipes away our whole economic system.

It's not fair to call OpenAI a cult, but when I asked several of the company's top brass if someone could comfortably work there if they didn't believe AGI was truly coming—and that its arrival would mark one of the greatest moments in human history—most executives didn't think so. *Why would a nonbeliever want to work here*? they wondered. The assumption is that the workforce—now at approximately 500, though it might have grown since you began reading this paragraph—has self-selected to include only the faithful. At the very least, as Altman puts it, once you get hired, it seems inevitable that you'll be drawn into the spell.

At the same time, OpenAI is not the company it once was. It was founded as a purely nonprofit research operation, but today most of its employees technically work for a profit-making entity that is reportedly valued at almost $30 billion. Altman and his team now face the pressure to deliver a revolution in every product cycle, in a way that satisfies the commercial demands of investors and keeps ahead in a fiercely competitive landscape. All while hewing to a quasi-messianic mission to elevate humanity rather than exterminate it.

That kind of pressure—not to mention the unforgiving attention of the entire world—can be a debilitating force. The Beatles set off colossal waves of cultural change, but they anchored their revolution for only so long: Six years after chiming that unforgettable chord they weren't even a band anymore. The maelstrom OpenAI has unleashed will almost certainly be far bigger. But the leaders of OpenAI swear they'll stay the course. All they want to do, they say, is build computers smart enough and safe enough to end history, thrusting humanity into an era of unimaginable bounty.

GROWING UP IN the late '80s and early '90s, Sam Altman was a nerdy kid who gobbled up science fiction and *Star Wars*. The worlds built by early sci-fi writers often had humans living with—or competing with—superintelligent AI systems. The idea of computers matching or exceeding human capabilities thrilled Altman, who had been coding since his fingers could barely cover a keyboard. When he was 8, his parents bought him a Macintosh LC II. One night he was up late playing with it and the thought popped into his head: "Someday this computer is going to learn to think." When he arrived at Stanford as an undergrad in 2003, he hoped to help make that happen and took courses in AI. But "it wasn't working at all," he'd later say. The field was still mired in an innovation trough known as AI winter. Altman dropped out to enter the startup world; his company Loopt was in the tiny first batch of wannabe organizations in Y Combinator, which would become the world's most famed incubator.

In February 2014, Paul Graham, YC's founding guru, chose then-28-year-old Altman to succeed him. "Sam is one of the smartest people I know," Graham wrote in the announcement, "and understands startups better than perhaps anyone I know, including myself." But Altman saw YC as something bigger than a launchpad for companies. "We are not about startups," he told me soon after taking over. "We are about innovation, because we believe that is how you make the future great for everyone." In Altman's view, the point of cashing in on all those unicorns was not to pack the partners' wallets but to fund species-level transformations. He began a research wing, hoping to fund ambitious projects to solve the world's biggest problems. But AI, in his mind, was the one realm of innovation to rule them all: a superintelligence that could address humanity's problems better than humanity could.

As luck would have it, Altman assumed his new job just as AI winter was turning into an abundant spring. Computers were now performing amazing feats, via deep learning and neural networks, like labeling photos, translating text, and optimizing sophisticated ad networks. The advances convinced him that for the first time, AGI was actually within reach. Leaving it in the hands of big corporations, however, worried him. He felt those companies would be too fixated on their products to seize the opportunity to develop AGI as soon as possible. And if they did create AGI, they might recklessly unleash it upon the world without the necessary precautions.
At the time, Altman had been thinking about running for governor of California. But he realized that he was perfectly positioned to do something bigger—to lead a company that would change humanity itself. “AGI was going to get built exactly once,” he told me in 2021. “And there were not that many people that could do a good job running OpenAI. I was lucky to have a set of experiences in my life that made me really positively set up for this.”

Altman began talking to people who might help him start a new kind of AI company, a nonprofit that would direct the field toward responsible AGI. One kindred spirit was Tesla and SpaceX CEO Elon Musk. As Musk would later tell CNBC, he had become concerned about AI’s impact after having some marathon discussions with Google cofounder Larry Page. Musk said he was dismayed that Page had little concern for safety and also seemed to regard the rights of robots as equal to humans. When Musk shared his concerns, Page accused him of being a “speciesist.” Musk also understood that, at the time, Google employed much of the world’s AI talent. He was willing to spend some money for an effort more amenable to Team Human.

Within a few months Altman had raised money from Musk (who pledged $100 million, and his time) and Reid Hoffman (who donated $10 million). Other funders included Peter Thiel, Jessica Livingston, Amazon Web Services, and YC Research. Altman began to stealthily recruit a team. He limited the search to AGI believers, a constraint that narrowed his options but one he considered critical. “Back in 2015, when we were recruiting, it was almost considered a career killer for an AI researcher to say that you took AGI seriously,” he says. “But I wanted people who took it seriously.”

Greg Brockman is now OpenAI's president. PHOTOGRAPH: JESSICA CHOU

Greg Brockman, the chief technology officer of Stripe, was one such person, and he agreed to be OpenAI's CTO. Another key cofounder would be Andrej Karpathy, who had been at Google Brain, the search giant's cutting-edge AI research operation. But perhaps Altman's most sought-after target was a Russian-born engineer named Ilya Sutskever.

Sutskever's pedigree was unassailable. His family had emigrated from Russia to Israel, then to Canada. At the University of Toronto he had been a standout student under Geoffrey Hinton, known as the godfather of modern AI for his work on deep learning and neural networks. Hinton, who is still close to Sutskever, marvels at his protégé's wizardry. Early in Sutskever's tenure at the lab, Hinton had given him a complicated project. Sutskever got tired of writing code to do the requisite calculations, and he told Hinton it would be easier if he wrote a custom programming language for the task. Hinton got a bit annoyed and tried to warn his student away from what he assumed would be a monthlong distraction. Then Sutskever came clean: "I did it this morning."

Sutskever became an AI superstar, coauthoring a breakthrough paper that showed how AI could learn to recognize images simply by being exposed to huge volumes of data. He ended up, happily, as a key scientist on the Google Brain team.

In mid-2015 Altman cold-emailed Sutskever to invite him to dinner with Musk, Brockman, and others at the swank Rosewood Hotel on Palo Alto's Sand Hill Road. Only later did Sutskever figure out that he was the guest of honor. "It was kind of a general conversation about AI and AGI in the future," he says. More specifically, they discussed "whether Google and DeepMind were so far ahead that it would be impossible to catch up to them, or whether it was still possible to, as Elon put it, create a lab which would be a counterbalance." While no one at the dinner explicitly tried to recruit Sutskever, the conversation hooked him.

Sutskever wrote an email to Altman soon after, saying he was game to lead the project—but the message got stuck in his drafts folder. Altman circled back, and after months fending off Google's counteroffers, Sutskever signed on. He would soon become the soul of the company and its driving force in research.

Sutskever joined Altman and Musk in recruiting people to the project, culminating in a Napa Valley retreat where several prospective OpenAI researchers fueled each other's excitement. Of course, some targets would resist the lure. John Carmack, the legendary gaming coder behind *Doom*, *Quake*, and countless other titles, declined an Altman pitch.

OpenAI officially launched in December 2015. At the time, when I interviewed Musk and Altman, they presented the project to me as an effort to make AI safe and accessible by sharing it with the world. In other words, open source. OpenAI, they told me, was not going to apply for patents. Everyone could make use of their breakthroughs. Wouldn't that be empowering some future Dr. Evil? I wondered. Musk said that was a good question. But Altman had an answer: Humans are generally good, and because OpenAI would provide powerful tools for that vast majority, the bad actors would be overwhelmed. He admitted that if Dr. Evil were to use the tools to build something that couldn't be counteracted, "then we're in a really bad place." But both Musk and Altman believed that the safer course for AI would be in the hands of a research operation not polluted by the profit motive, a persistent temptation to ignore the needs of humans in the search for boffo quarterly results.

Altman cautioned me not to expect results soon. “This is going to look like a research lab for a long time,” he said.

There was another reason to tamp down expectations. Google and the others had been developing and applying AI for years. While OpenAI had a billion dollars committed (largely via Musk), an ace team of researchers and engineers, and a lofty mission, it had no clue about how to pursue its goals. Altman remembers a moment when the small team gathered in Brockman’s apartment—they didn’t have an office yet. “I was like, what should we do?”

> Altman remembers a moment when the small team gathered in Brockman’s apartment—they didn’t have an office yet. “I was like, what should we do?”

I had breakfast in San Francisco with Brockman a little more than a year after OpenAI’s founding. For the CTO of a company with the word *open* in its name, he was pretty parsimonious with details. He did affirm that the nonprofit could afford to draw on its initial billion-dollar donation for a while. The salaries of the 25 people on its staff—who were being paid at far less than market value—ate up the bulk of OpenAI’s expenses. “The goal for us, the thing that we’re really pushing on,” he said, “is to have the systems that can do things that humans were just not capable of doing before.” But for the time being, what that looked like was a bunch of researchers publishing papers. After the interview, I walked him to the company’s newish office in the Mission District, but he allowed me to go no further than the vestibule. He did duck into a closet to get me a T-shirt.

Had I gone in and asked around, I might have learned exactly how much OpenAI *was* floundering. Brockman now admits that “nothing was working.” Its researchers were tossing algorithmic spaghetti toward the ceiling to see what stuck. They delved into systems that solved video games and spent considerable effort on robotics. “We knew *what* we wanted to do,” says Altman. “We knew *why* we wanted to do it. But we had no idea *how*.”

But they *believed*. Supporting their optimism were the steady improvements in artificial neural networks that used deep-learning techniques."The general idea is, don't bet against deep learning," says Sutskever. Chasing AGI, he says, "wasn't totally crazy. It was only moderately crazy."

OpenAI's road to relevance really started with its hire of an as-yet-unheralded researcher named Alec Radford, who joined in 2016, leaving the small Boston AI company he'd cofounded in his dorm room. After accepting OpenAI's offer, he told his high school alumni magazine that taking this new role was "kind of similar to joining a graduate program"—an open-ended, low-pressure perch to research AI.

The role he would actually play was more like Larry Page inventing PageRank.

Radford, who is press-shy and hasn't given interviews on his work, responds to my questions about his early days at OpenAI via a long email exchange. His biggest interest was in getting neural nets to interact with humans in lucid conversation. This was a departure from the traditional scripted model of making a chatbot, an approach used in everything from the primitive ELIZA to the popular assistants Siri and Alexa—all of which kind of sucked. "The goal was to see if there was any task, any setting, any domain, any *anything* that language models could be useful for," he writes. At the time, he explains, "language models were seen as novelty toys that could only generate a sentence that made sense once in a while, and only then if you really squinted." His first experiment involved scanning 2 billion Reddit comments to train a language model. Like a lot of OpenAI's early experiments, it flopped. No matter. The 23-year-old had permission to keep going, to fail again. "We were just like, Alec is great, let him do his thing," says Brockman.
His next major experiment was shaped by OpenAI's limitations of computer power, a constraint that led him to experiment on a smaller data set that focused on a single domain—Amazon product reviews. A researcher had gathered about 100 million of those. Radford trained a language model to simply predict the next character in generating a user review.

> Radford began experimenting with the transformer architecture. "I made more progress in two weeks than I did over the past two years," he says.

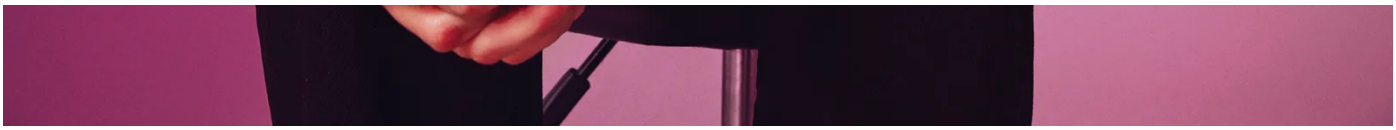
But then, on its own, the model figured out whether a review was positive or negative—and when you programmed the model to create something positive or negative, it delivered a review that was adulatory or scathing, as requested. (The prose was admittedly clunky: "I love this weapons look ... A must watch for any man who love Chess!") "It was a complete surprise," Radford says. The sentiment of a review—its favorable or disfavorable gist—is a complex function of semantics, but somehow a part of Radford's system had gotten a feel for it. Within OpenAI, this part of the neural net came to be known as the "unsupervised sentiment neuron."

Sutskever and others encouraged Radford to expand his experiments beyond Amazon reviews, to use his insights to train neural nets to converse or answer questions on a broad range of subjects.

And then good fortune smiled on OpenAI. In early 2017, an unheralded preprint of a research paper appeared, coauthored by eight Google researchers. Its official title was "Attention Is All You Need," but it came to be known as the "transformer paper," named so both to reflect the game-changing nature of the idea and to honor the toys that transmogrified from trucks to giant robots. Transformers made it possible for a neural net to understand—and generate—language much more efficiently. They did this by analyzing chunks of prose in parallel and figuring out which elements merited "attention." This hugely optimized the process of generating coherent text to respond to prompts. Eventually, people came to realize that the same technique could also generate images and even video. Though the transformer paper would become known as the catalyst for the current AI frenzy—think of it as the Elvis that made the Beatles possible—at the time Ilya Sutskever was one of only a handful of people who understood how powerful the breakthrough was. "The real *aha* moment was when Ilya saw the transformer come out," Brockman says. "He was like, 'That's what we've been waiting for.' That's been our strategy—to push hard on problems and then have faith that we or someone in the field will manage to figure out the missing ingredient."

Radford began experimenting with the transformer architecture. "I made more progress in two weeks than I did over the past two years," he says. He came to understand that the key to getting the most out of the new model was to add scale—to train it on fantastically large data sets. The idea was dubbed "Big Transformer" by Radford's collaborator Rewon Child.

This approach required a change of culture at OpenAI and a focus it had previously lacked. “In order to take advantage of the transformer, you needed to scale it up,” says Adam D’Angelo, the CEO of Quora, who sits on OpenAI’s board of directors. “You need to run it more like an engineering organization. You can’t have every researcher trying to do their own thing and training their own model and make elegant things that you can publish papers on. You have to do this more tedious, less elegant work.” That, he added, was something OpenAI was able to do, and something no one else did.



Mira Murati, OpenAI's chief technology officer. PHOTOGRAPH: JESSICA CHOU

The name that Radford and his collaborators gave the model they created was an acronym for "generatively pretrained transformer"—GPT-1. Eventually, this model came to be generically known as "generative AI." To build it, they drew on a collection of 7,000 unpublished books, many in the genres of romance, fantasy, and adventure, and refined it on Quora questions and answers, as well as thousands of passages taken from middle school and high school exams. All in all, the model included 117 million parameters, or variables. And it outperformed everything that had come before in understanding language and generating answers. But the most dramatic result was that processing such a massive amount of data allowed the model to offer up results *beyond* its training, providing expertise in brand-new domains. These unplanned robot capabilities are called zero-shots. They still baffle researchers—and account for the queasiness that many in the field have about these so-called large language models.

Radford remembers one late night at OpenAI's office. "I just kept saying over and over, 'Well, that's cool, but I'm pretty sure it won't be able to do *x*.' And then I would quickly code up an evaluation and, sure enough, it could kind of do *x*."

Each GPT iteration would do better, in part because each one gobbled an order of magnitude more data than the previous model. Only a year after creating the first iteration, OpenAI trained GPT-2 on the open internet with an astounding 1.5 billion parameters. Like a toddler mastering speech, its responses got better and more coherent. So much so that OpenAI hesitated to release the program into the wild. Radford was worried that it might be used to generate spam. "I remember reading Neal Stephenson's *Anathem* in 2008, and in that book the internet was overrun with spam generators," he says. "I had thought that was really far-fetched, but as I worked on language models over the years and they got better, the uncomfortable realization that it was a real possibility set in."

In fact, the team at OpenAI was starting to think it wasn't such a good idea after all to put its work where Dr. Evil could easily access it. "We thought that open-sourcing

GPT-2 could be really dangerous," says chief technology officer Mira Murati, who started at the company in 2018. "We did a lot of work with misinformation experts and did some red-teaming. There was a lot of discussion internally on how much to release." Ultimately, OpenAI temporarily withheld the full version, making a less powerful version available to the public. When the company finally shared the full version, the world managed just fine—but there was no guarantee that more powerful models would avoid catastrophe.

The very fact that OpenAI was making products smart enough to be deemed dangerous, and was grappling with ways to make them safe, was proof that the company had gotten its mojo working. "We'd figured out the formula for progress, the formula everyone perceives now—the oxygen and the hydrogen of deep learning is computation with a large neural network and data," says Sutskever.

To Altman, it was a mind-bending experience. "If you asked the 10-year-old version of me, who used to spend a lot of time daydreaming about AI, what was going to happen, my pretty confident prediction would have been that first we're gonna have robots, and they're going to perform all physical labor. Then we're going to have systems that can do basic cognitive labor. A really long way after that, maybe we'll have systems that can do complex stuff like proving mathematical theorems. Finally we will have AI that can create new things and make art and write and do these deeply human things. That was a terrible prediction—it's going exactly the other direction."

The world didn't know it yet, but Altman and Musk's research lab had begun a climb that plausibly creeps toward the summit of AGI. The crazy idea behind OpenAI suddenly was not so crazy.

BY EARLY 2018, OpenAI was starting to focus productively on large language models, or LLMs. But Elon Musk wasn't happy. He felt that the progress was insufficient—or maybe he felt that now that OpenAI was on to something, it needed leadership to seize its advantage. Or maybe, as he'd later explain, he felt that safety should be more of a priority. Whatever his problem was, he had a solution: Turn everything over to him. He proposed taking a majority stake in the company, adding it to the portfolio of his multiple full-time jobs (Tesla, SpaceX) and supervisory obligations (Neuralink and the Boring Company).

Musk believed he had a *right* to own OpenAI. “It wouldn’t exist without me,” he later told CNBC. “I came up with the name!” (True.) But Altman and the rest of OpenAI’s brain trust had no interest in becoming part of the Muskiverse. When they made this clear, Musk cut ties, providing the public with the incomplete explanation that he was leaving the board to avoid a conflict with Tesla’s AI effort. His farewell came at an all-hands meeting early that year where he predicted that OpenAI would fail. And he called at least one of the researchers a “jackass.”

He also took his money with him. Since the company had no revenue, this was an existential crisis. “Elon is cutting off his support,” Altman said in a panicky call to Reid Hoffman. “What do we do?” Hoffman volunteered to keep the company afloat, paying overhead and salaries.

But this was a temporary fix; OpenAI had to find big bucks elsewhere. Silicon Valley loves to throw money at talented people working on trendy tech. But not so much if they are working at a nonprofit. It had been a massive lift for OpenAI to get its first billion. To train and test new generations of GPT—and then access the computation it takes to deploy them—the company needed another billion, and fast. And that would only be the start.

> Somewhere in the restructuring documents is a clause to the effect that, if the company does manage to create AGI, all financial arrangements will be reconsidered. After all, it will be a new world from that point on.

So in March 2019, OpenAI came up with a bizarre hack. It would remain a nonprofit, fully devoted to its mission. But it would also create a for-profit entity. The actual structure of the arrangement is hopelessly baroque, but basically the entire company is now engaged in a “capped” profitable business. If the cap is reached—the number isn’t public, but its own charter, if you read between the lines, suggests it might be in the trillions—everything beyond that reverts to the nonprofit research lab. The novel scheme was almost a quantum approach to incorporation: Behold a company that, depending on your time-space point of view, is for-profit and nonprofit. The details

are embodied in charts full of boxes and arrows, like the ones in the middle of a scientific paper where only PhDs or dropout geniuses dare to tread. When I suggest to Sutskever that it looks like something the as-yet-unconceived GPT-6 might come up with if you prompted it for a tax dodge, he doesn't warm to my metaphor. "It's not about accounting," he says.

But accounting is critical. A for-profit company optimizes for, well, profits. There's a reason why companies like Meta feel pressure from shareholders when they devote billions to R&D. How could this not affect the way a firm operates? And wasn't avoiding commercialism the reason why Altman made OpenAI a nonprofit to begin with? According to COO Brad Lightcap, the view of the company's leaders is that the board, which is still part of the nonprofit controlling entity, will make sure that the drive for revenue and profits won't overwhelm the original idea. "We needed to maintain the mission as the reason for our existence," he says, "It shouldn't just be in spirit, but encoded in the structure of the company." Board member Adam D'Angelo says he takes this responsibility seriously: "It's my job, along with the rest of the board, to make sure that OpenAI stays true to its mission."

Potential investors were warned about those boundaries, Lightcap explains. "We have a legal disclaimer that says you, as an investor, stand to lose all your money," he says. "We are not here to make your return. We're here to achieve a technical mission, foremost. And, oh, by the way, we don't really know what role money will play in a post-AGI world."

That last sentence is not a throwaway joke. OpenAI's plan really does include a reset in case computers reach the final frontier. Somewhere in the restructuring documents is a clause to the effect that, if the company does manage to create AGI, all financial arrangements will be reconsidered. After all, it will be a new world from that point on. Humanity will have an alien partner that can do much of what we do, only better. So previous arrangements might effectively be kaput.

There is, however, a hitch: At the moment, OpenAI doesn't claim to know what AGI really *is*. The determination would come from the board, but it's not clear how the board would define it. When I ask Altman, who is on the board, for clarity, his response is anything but open. "It's not a single Turing test, but a number of things we might use," he says. "I would happily tell you, but I like to keep confidential

conversations private. I realize that is unsatisfyingly vague. But we don't know what it's going to be like at that point."

Nonetheless, the inclusion of the "financial arrangements" clause isn't just for fun: OpenAI's leaders think that if the company is successful enough to reach its lofty profit cap, its products will probably have performed well enough to reach AGI. Whatever that is.

"My regret is that we've chosen to double down on the term AGI," Sutskever says. "In hindsight it is a confusing term, because it emphasizes generality above all else. GPT-3 is general AI, but yet we don't really feel comfortable calling it AGI, because we want human-level competence. But back then, at the beginning, the idea of OpenAI was that superintelligence is attainable. It is the endgame, the final purpose of the field of AI."

Those caveats didn't stop some of the smartest venture capitalists from throwing money at OpenAI during its 2019 funding round. At that point, the first VC firm to invest was Khosla Ventures, which kicked in $50 million. According to Vinod Khosla, it was double the size of his largest initial investment. "If we lose, we lose 50 million bucks," he says. "If we win, we win 5 billion." Others investors reportedly would include elite VC firms Thrive Capital, Andreessen Horowitz, Founders Fund, and Sequoia.

The shift also allowed OpenAI's employees to claim some equity. But not Altman. He says that originally he intended to include himself but didn't get around to it. Then he decided that he didn't need any piece of the $30 billion company that he'd cofounded and leads. "Meaningful work is more important to me," he says. "I don't think about it. I honestly don't get why people care so much."

Because ... not taking a stake in the company you cofounded is weird?

"If I didn't already have a ton of money, it would be much weirder," he says. "It does seem like people have a hard time imagining ever having enough money. But I feel like I have enough." (Note: For Silicon Valley, this is *extremely* weird.) Altman joked that he's considering taking one share of equity "so I never have to answer that question again."

Ilya Sutskever, OpenAI's chief scientist. PHOTOGRAPH: JESSICA CHOU

THE BILLION-DOLLAR VC round wasn't even table stakes to pursue OpenAI's vision. The miraculous Big Transformer approach to creating LLMs required Big Hardware. Each iteration of the GPT family would need exponentially more power—GPT-2 had over a billion parameters, and GPT-3 would use 175 billion. OpenAI was now like Quint in *Jaws* after the shark hunter sees the size of the great white. "It turned out we didn't know how much of a bigger boat we needed," Altman says.

Obviously, only a few companies in existence had the kind of resources OpenAI required. "We pretty quickly zeroed in on Microsoft," says Altman. To the credit of Microsoft CEO Satya Nadella and CTO Kevin Scott, the software giant was able to get over an uncomfortable reality: After more than 20 years and billions of dollars spent on a research division with supposedly cutting-edge AI, the Softies needed an innovation infusion from a tiny company that was only a few years old. Scott says that it wasn't just Microsoft that fell short—"it was everyone." OpenAI's focus on pursuing AGI, he says, allowed it to accomplish a moonshot-ish achievement that the heavy hitters weren't even aiming for. It also proved that not pursuing generative

AI was a lapse that Microsoft needed to address. “One thing you just very clearly need is a frontier model,” says Scott.

Microsoft originally chipped in a billion dollars, paid off in computation time on its servers. But as both sides grew more confident, the deal expanded. Microsoft now has sunk $13 billion into OpenAI. (“Being on the frontier is a very expensive proposition,” Scott says.)

Of course, because OpenAI couldn’t exist without the backing of a huge cloud provider, Microsoft was able to cut a great deal for itself. The corporation bargained for what Nadella calls “non-controlling equity interest” in OpenAI’s for-profit side—reportedly 49 percent. Under the terms of the deal, some of OpenAI’s original ideals of granting equal access to all were seemingly dragged to the trash icon. (Altman objects to this characterization.) Now, Microsoft has an exclusive license to commercialize OpenAI’s tech. And OpenAI also has committed to use Microsoft’s cloud exclusively. In other words, without even taking its cut of OpenAI’s profits (reportedly Microsoft gets 75 percent until its investment is paid back), Microsoft gets to lock in one of the world’s most desirable new customers for its Azure web services. With those rewards in sight, Microsoft wasn’t even bothered by the clause that demands reconsideration if OpenAI achieves general artificial intelligence, whatever that is. “At that point,” says Nadella, “all bets are off.” It might be the last invention of humanity, he notes, so we might have bigger issues to consider once machines are smarter than we are.

By the time Microsoft began unloading Brinks trucks’ worth of cash into OpenAI ($2 billion in 2021, and the other $10 billion earlier this year), OpenAI had completed GPT-3, which, of course, was even more impressive than its predecessors. When Nadella saw what GPT-3 could do, he says, it was the first time he deeply understood that Microsoft had snared something truly transformative. “We started observing all those emergent properties.” For instance, GPT had taught itself how to program computers. “We didn’t train it on coding—it just got good at coding!” he says. Leveraging its ownership of GitHub, Microsoft released a product called Copilot that uses GPT to churn out code literally on command. Microsoft would later integrate OpenAI technology in new versions of its workplace products. Users pay a premium for those, and a cut of that revenue gets logged to OpenAI’s ledger.

Some observers professed whiplash at OpenAI's one-two punch: creating a for-profit component and reaching an exclusive deal with Microsoft. How did a company that promised to remain patent-free, open source, and totally transparent wind up giving an exclusive license of its tech to the world's biggest software company? Elon Musk's remarks were particularly lacerating. "This does seem like the opposite of open—OpenAI is essentially captured by Microsoft," he posted on Twitter. On CNBC, he elaborated with an analogy: "Let's say you founded an organization to save the Amazon rainforest, and instead you became a lumber company, chopped down the forest, and sold it."

Musk's jibes might be dismissed as bitterness from a rejected suitor, but he wasn't alone. "The whole vision of it morphing the way it did feels kind of gross," says John Carmack. (He does specify that he's still excited about the company's work.) Another prominent industry insider, who prefers to speak without attribution, says, "OpenAI has turned from a small, somewhat open research outfit into a secretive product-development house with an unwarranted superiority complex."

Even some employees had been turned off by OpenAI's venture into the for-profit world. In 2019, several key executives, including head of research Dario Amodei, left to start a rival AI company called Anthropic. They recently told *The New York Times* that OpenAI had gotten too commercial and had fallen victim to mission drift.

Another OpenAI defector was Rewon Child, a main technical contributor to the GPT-2 and GPT-3 projects. He left in late 2021 and is now at Inflection AI, a company led by former DeepMind cofounder Mustafa Suleyman.

Altman professes not to be bothered by defections, dismissing them as simply the way Silicon Valley works. "Some people will want to do great work somewhere else, and that pushes society forward," he says. "That absolutely fits our mission."

UNTIL NOVEMBER OF last year, awareness of OpenAI was largely confined to people following technology and software development. But as the whole world now knows, OpenAI took the dramatic step of releasing a consumer product late that month, built on what was then the most recent iteration of GPT, version 3.5. For months, the company had been internally using a version of GPT with a conversational interface. It was especially important for what the company called

"truth-seeking." That means that via dialog, the user could coax the model to provide responses that would be more trustworthy and complete. ChatGPT, optimized for the masses, could allow anyone to instantly tap into what seemed to be an endless source of knowledge simply by typing in a prompt—and then continue the conversation as if hanging out with a fellow human who just happened to know everything, albeit one with a penchant for fabrication.

Within OpenAI, there was a lot of debate about the wisdom of releasing a tool with such unprecedented power. But Altman was all for it. The release, he explains, was part of a strategy designed to acclimate the public to the reality that artificial intelligence is destined to change their everyday lives, presumably for the better. Internally, this is known as the "iterative deployment hypothesis." Sure, ChatGPT would create a stir, the thinking went. After all, here was something anyone could use that was smart enough to get college-level scores on the SATs, write a B-minus essay, and summarize a book within seconds. You could ask it to write your funding proposal or summarize a meeting and then request it to do a rewrite in Lithuanian or as a Shakespeare sonnet or in the voice of someone obsessed with toy trains. In a few seconds, pow, the LLM would comply. Bonkers. But OpenAI saw it as a table-setter for its newer, more coherent, more capable, and scarier successor, GPT-4, trained with a reported 1.7 trillion parameters. (OpenAI won't confirm the number, nor will it reveal the data sets.)

Altman explains why OpenAI released ChatGPT when GPT-4 was close to completion, undergoing safety work. "With ChatGPT, we could introduce chatting but with a much less powerful backend, and give people a more gradual adaptation," he says. "GPT-4 was a lot to get used to at once." By the time the ChatGPT excitement cooled down, the thinking went, people might be ready for GPT-4, which can pass the bar exam, plan a course syllabus, and write a book within seconds. (Publishing houses that produced genre fiction were indeed flooded with AI-generated bodice rippers and space operas.)

A cynic might say that a steady cadence of new products is tied to the company's commitment to investors, and equity-holding employees, to make some money. OpenAI now charges customers who use its products frequently. But OpenAI insists that its true strategy is to provide a soft landing for the singularity. "It doesn't make sense to just build AGI in secret and drop it on the world," Altman says. "Look back at

the industrial revolution—everyone agrees it was great for the world," says Sandhini Agarwal, an OpenAI policy researcher. "But the first 50 years were really painful. There was a lot of job loss, a lot of poverty, and then the world adapted. We're trying to think how we can make the period before adaptation of AGI as painless as possible."

Sutskever puts it another way: "You want to build larger and more powerful intelligences and keep them in your basement?"

Even so, OpenAI was stunned at the reaction to ChatGPT. "Our internal excitement was more focused on GPT-4," says Murati, the CTO. "And so we didn't think ChatGPT was really going to change everything." To the contrary, it galvanized the public to the reality that AI had to be dealt with, *now.* ChatGPT became the fastest-growing consumer software in history, amassing a reported 100 million users. (Not-so-OpenAI won't confirm this, saying only that it has "millions of users.") "I underappreciated how much making an easy-to-use conversational interface to an LLM would make it much more intuitive for everyone to use," says Radford.

ChatGPT was of course delightful and astonishingly useful, but also scary—prone to "hallucinations" of plausible but shamefully fabulist details when responding to prompts. Even as journalists wrung their hands about the implications, however, they effectively endorsed ChatGPT by extolling its powers.

The clamor got even louder in February when Microsoft, taking advantage of its multibillion-dollar partnership, released a ChatGPT-powered version of its search engine Bing. CEO Nadella was euphoric that he had beaten Google to the punch in introducing generative AI to Microsoft's products. He taunted the search king, which had been cautious in releasing its own LLM into products, to do the same. "I want people to know we made them dance," he said.

In so doing, Nadella triggered an arms race that tempted companies big and small to release AI products before they were fully vetted. He also a triggered a new round of media coverage that kept wider and wider circles of people up at night: interactions with Bing that unveiled the chatbot's shadow side, replete with unnerving professions of love, an envy of human freedom, and a weak resolve to withhold

misinformation. As well as an unseemly habit of creating hallucinatory misinformation of its own.

But if OpenAI's products were forcing people to confront the implications of artificial intelligence, Altman figured, so much the better. It was time for the bulk of humankind to come off the sidelines in discussions of how AI might affect the future of the species.

PHOTOGRAPH: JESSICA CHOU

OpenAI's San Francisco headquarters is unmarked; but inside, the coffee is awesome. PHOTOGRAPH: JESSICA CHOU

AS SOCIETY STARTED to prioritize thinking through all the potential drawbacks of AI—job loss, misinformation, human extinction—OpenAI set about placing itself in the center of the discussion. Because if regulators, legislators, and doomsayers mounted a charge to smother this nascent alien intelligence in its cloud-based cradle, OpenAI would be their chief target anyway. "Given our current visibility, when things go wrong, even if those things were built by a different company, that's still a problem for us, because we're viewed as the face of this technology right now," says Anna Makanju, OpenAI's chief policy officer.

Makanju is a Russian-born DC insider who served in foreign policy roles at the US Mission to the United Nations, the US National Security Council, and the Defense Department, and in the office of Joe Biden when he was vice president. "I have lots of preexisting relationships, both in the US government and in various European governments," she says. She joined OpenAI in September 2021. At the time, very few people in government gave a hoot about generative AI. Knowing that OpenAI's products would soon change that, she began to introduce Altman to administration

officials and legislators, making sure that they'd hear the good news and the bad from OpenAI first.

"Sam has been extremely helpful, but also very savvy, in the way that he has dealt with members of Congress," says Richard Blumenthal, the chair of the Senate Judiciary Committee. He contrasts Altman's behavior with that of the younger Bill Gates, who unwisely stonewalled legislators when Microsoft was under antitrust investigations in the 1990s. "Altman, by contrast, was happy to spend an hour or more sitting with me to try to educate me," says Blumenthal. "He didn't come with an army of lobbyists or minders. He demonstrated ChatGPT. It was mind-blowing."

In Blumenthal, Altman wound up making a semi-ally of a potential foe. "Yes," the senator admits. "I'm excited about both the upside and the potential perils." OpenAI didn't shrug off discussion of those perils, but presented itself as the force best positioned to mitigate them. "We had 100-page system cards on all the red-teaming safety valuations," says Makanju. (Whatever that meant, it didn't stop users and journalists from endlessly discovering ways to jailbreak the system.)

By the time Altman made his first appearance in a congressional hearing—fighting a fierce migraine headache—the path was clear for him to sail through in a way that Bill Gates or Mark Zuckerberg could never hope to. He faced almost none of the tough questions and arrogant badgering that tech CEOs now routinely endure after taking the oath. Instead, senators asked Altman for advice on how to regulate AI, a pursuit Altman enthusiastically endorsed.

The paradox is that no matter how assiduously companies like OpenAI red-team their products to mitigate misbehavior like deepfakes, misinformation efforts, and criminal spam, future models might get smart enough to foil the efforts of the measly minded humans who invented the technology yet are still naive enough to believe they can control it. On the other hand, if they go *too* far in making their models safe, it might hobble the products, making them less useful. One study indicated that more recent versions of GPT, which have improved safety features, are actually dumber than previous versions, making errors in basic math problems that earlier programs had aced. (Altman says that OpenAI's data doesn't confirm this. "Wasn't that study retracted?" he asks. No.)

It makes sense that Altman positions himself as a fan of regulation; after all, his mission is AGI, but safely. Critics have charged that he's gaming the process so that regulations would thwart smaller startups and give an advantage to OpenAI and other big players. Altman denies this. While he has endorsed, in principle, the idea of an international agency overseeing AI, he does feel that some proposed rules, like banning all copyrighted material from data sets, present unfair obstacles. He pointedly didn't sign a widely distributed letter urging a six-month moratorium on developing more powerful AI systems. But he and other OpenAI leaders did add their names to a one-sentence statement: "Mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war." Altman explains: "I said, 'Yeah, I agree with that. One-minute discussion."

As one prominent Silicon Valley founder notes, "It's rare that an industry raises their hand and says, 'We are going to be the end of humanity'—and then continues to work on the product with glee and alacrity."

OpenAI rejects this criticism. Altman and his team say that working and releasing cutting-edge products is *the way* to address societal risks. Only by analyzing the responses to millions of prompts by users of ChatGPT and GPT-4 could they get the knowledge to ethically align their future products.

Still, as the company takes on more tasks and devotes more energy to commercial activities, some question how closely OpenAI can concentrate on the mission—especially the "mitigating risk of extinction" side. "If you think about it, they're actually building *five* businesses," says an AI industry executive, ticking them off with his fingers. "There's the product itself, the enterprise relationship with Microsoft, the developer ecosystem, and an app store. And, oh yes—they are also obviously doing an AGI research mission." Having used all five fingers, he recycles his index finger to add a sixth. "And of course, they're also doing the investment fund," he says, referring to a $175 million project to seed startups that want to tap into OpenAI technology. "These are different cultures, and in fact they're conflicting with a research mission."

I repeatedly asked OpenAI's execs how donning the skin of a product company has affected its culture. Without fail they insist that, despite the for-profit restructuring, despite the competition with Google, Meta, and countless startups, the mission is still

central. Yet OpenAI *has* changed. The nonprofit board might technically be in charge, but virtually everyone in the company is on the for-profit ledger. Its workforce includes lawyers, marketers, policy experts, and user-interface designers. OpenAI contracts with hundreds of content moderators to educate its models on inappropriate or harmful answers to the prompts offered by many millions of users. It's got product managers and engineers working constantly on updates to its products, and every couple of weeks it seems to ping reporters with demonstrations—just like other product-oriented Big Tech companies. Its offices look like an *Architectural Digest* spread. I have visited virtually every major tech company in Silicon Valley and beyond, and not one surpasses the coffee options in the lobby of OpenAI's headquarters in San Francisco.

Not to mention: It's obvious that the "openness" embodied in the company's name has shifted from the radical transparency suggested at launch. When I bring this up to Sutskever, he shrugs. "Evidently, times have changed," he says. But, he cautions, that doesn't mean that the prize is not the same. "You've got a technological transformation of such gargantuan, cataclysmic magnitude that, even if we all do our part, success is not guaranteed. But if it all works out we can have quite the incredible life."

> "The biggest thing we're missing is coming up with new ideas," says Brockman. "It's nice to have something that could be a virtual assistant. But that's not the dream. The dream is to help us solve problems we can't."

"I can't emphasize this enough—we didn't have a master plan," says Altman. "It was like we were turning each corner and shining a flashlight. We were willing to go through the maze to get to the end." Though the maze got twisty, the goal has not changed. "We still have our core mission—believing that safe AGI was this critically important thing that the world was not taking seriously enough."

Meanwhile, OpenAI is apparently taking its time to develop the next version of its large language model. It's hard to believe, but the company insists it has yet to begin

working on GPT-5, a product that people are, depending on point of view, either salivating about or dreading. Apparently, OpenAI is grappling with what an exponentially powerful improvement on its current technology actually looks like. "The biggest thing we're missing is coming up with new ideas," says Brockman. "It's nice to have something that could be a virtual assistant. But that's not the dream. The dream is to help us solve problems we can't."

Considering OpenAI's history, that next big set of innovations might have to wait until there's another breakthrough as major as transformers. Altman hopes that will come from OpenAI—"We want to be the best research lab in the world," he says—but even if not, his company will make use of others' advances, as it did with Google's work. "A lot of people around the world are going to do important work," he says.

It would also help if generative AI didn't create so many new problems of its own. For instance, LLMs need to be trained on huge data sets; clearly the most powerful ones would gobble up the whole internet. This doesn't sit well with some creators, and just plain people, who unwittingly provide content for those data sets and wind up somehow contributing to the output of ChatGPT. Tom Rubin, an elite intellectual property lawyer who officially joined OpenAI in March, is optimistic that the company will eventually find a balance that satisfies both its own needs and that of creators—including the ones, like comedian Sarah Silverman, who are suing OpenAI for using their content to train its models. One hint of OpenAI's path: partnerships with news and photo agencies like the Associated Press and Shutterstock to provide content for its models without questions of who owns what.

As I interview Rubin, my very human mind, subject to distractions you never see in LLMs, drifts to the arc of this company that in eight short years has gone from a floundering bunch of researchers to a Promethean behemoth that has changed the world. Its very success has led it to transform itself from a novel effort to achieve a scientific goal to something that resembles a standard Silicon Valley unicorn on its way to elbowing into the pantheon of Big Tech companies that affect our everyday lives. And here I am, talking with one of its key hires—a lawyer—not about neural net weights or computer infrastructure but copyright and fair use. Has this IP expert, I wonder, signed on to the mission, like the superintelligence-seeking voyagers who drove the company originally?

Rubin is nonplussed when I ask him whether he believes, as an article of faith, that AGI will happen and if he's hungry to make it so. "I can't even answer that," he says after a pause. When pressed further, he clarifies that, as an intellectual property lawyer, speeding the path to scarily intelligent computers is not his job. "From my perch, I look forward to it," he finally says.

Updated 9-7-23, 5:30pm EST: This story was updated to clarify Rewon Child's role at OpenAI, and the aim of a letter calling for a six-month pause on the most powerful AI models.

Styling by Turner/The Wall Group. Hair and Makeup by Hiroko Claus.

This article appears in the October 2023 issue. Subscribe now.

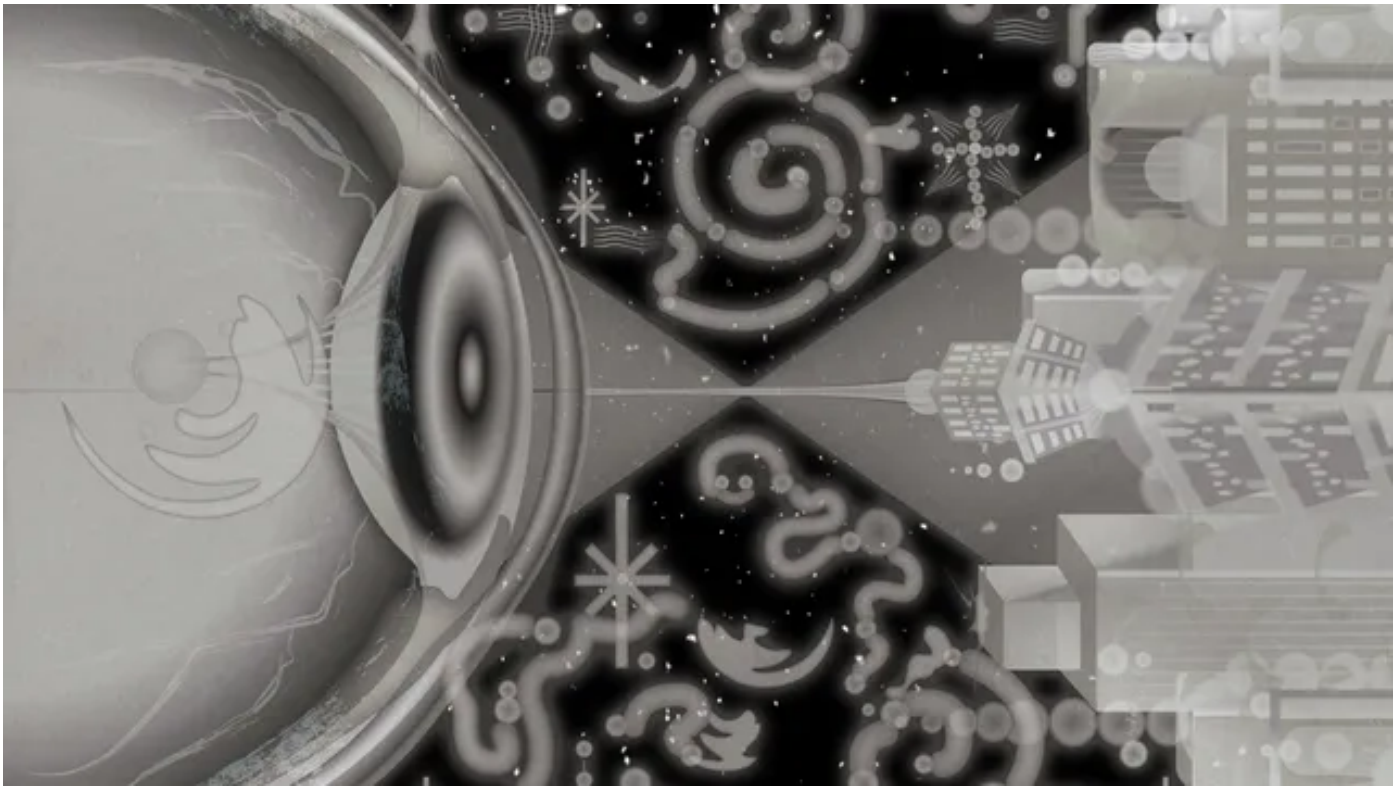
Let us know what you think about this article. Submit a letter to the editor at mail@wired.com.

Steven Levy covers the gamut of tech subjects for WIRED, in print and online, and has been contributing to the magazine since its inception. His weekly column, **Plaintext**, is exclusive to subscribers online but the newsletter version is open to all—**sign up here**. He has been writing about... Read more

EDITOR AT LARGE

TOPICS COVER STORY LONGREADS OPENAI ARTIFICIAL INTELLIGENCE MAGAZINE-31.10

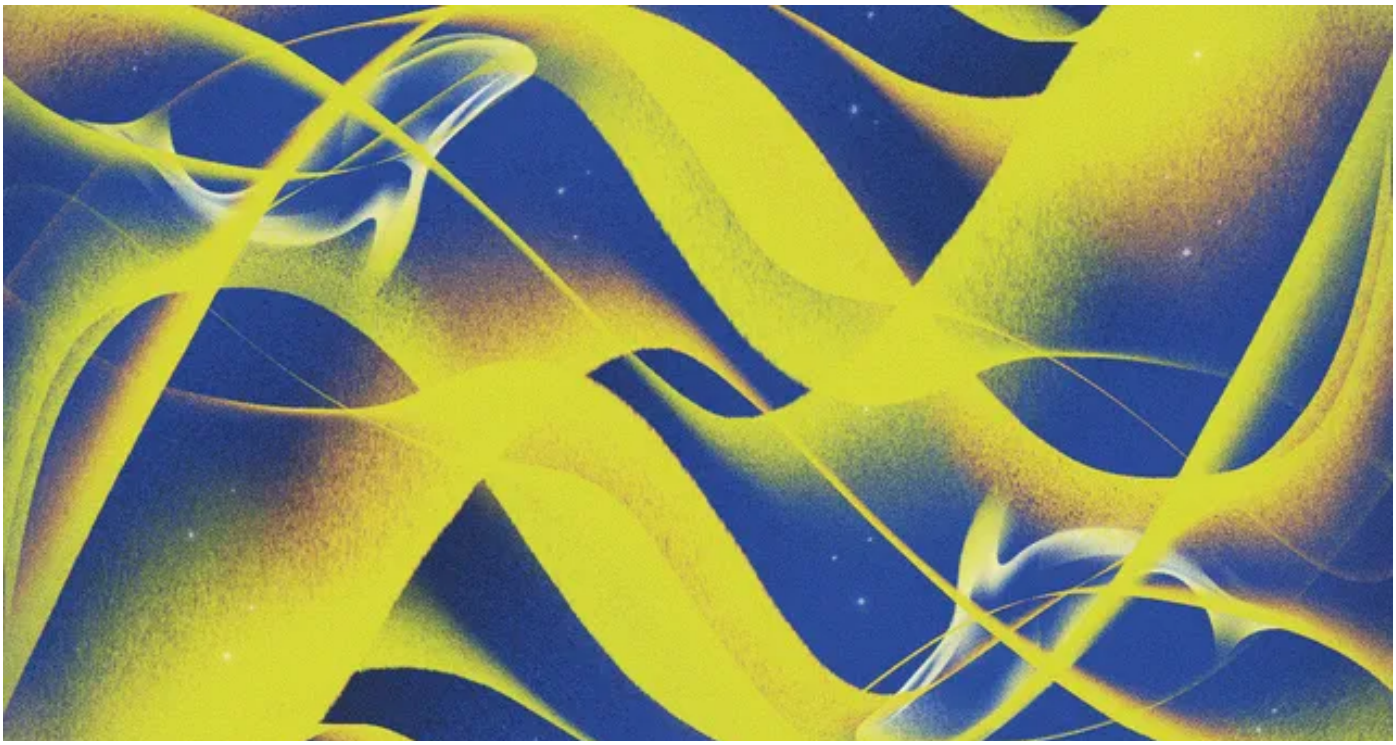
MORE FROM WIRED



The World Is Going Blind. Taiwan Offers a Warning, and a Cure

So many people are nearsighted on the island nation that they have already glimpsed what could be coming for the rest of us.

AMIT KATWALA



How to Use AI to Talk to Whales—and Save Life on Earth

With ecosystems in crisis, engineers and scientists are teaming up to decipher what animals are saying. Their hope: By truly listening to nature, humans will decide to protect it.

CAMILLE BROMLEY

Unhinged Conspiracies, AI Doppelgangers, and the Fractured Reality of Naomi Klein

When the internet confused her with a Covid truther, the author journeyed down the rabbit hole to understand why.

KATE KNIBBS

The Dark History *Oppenheimer* Didn't Show

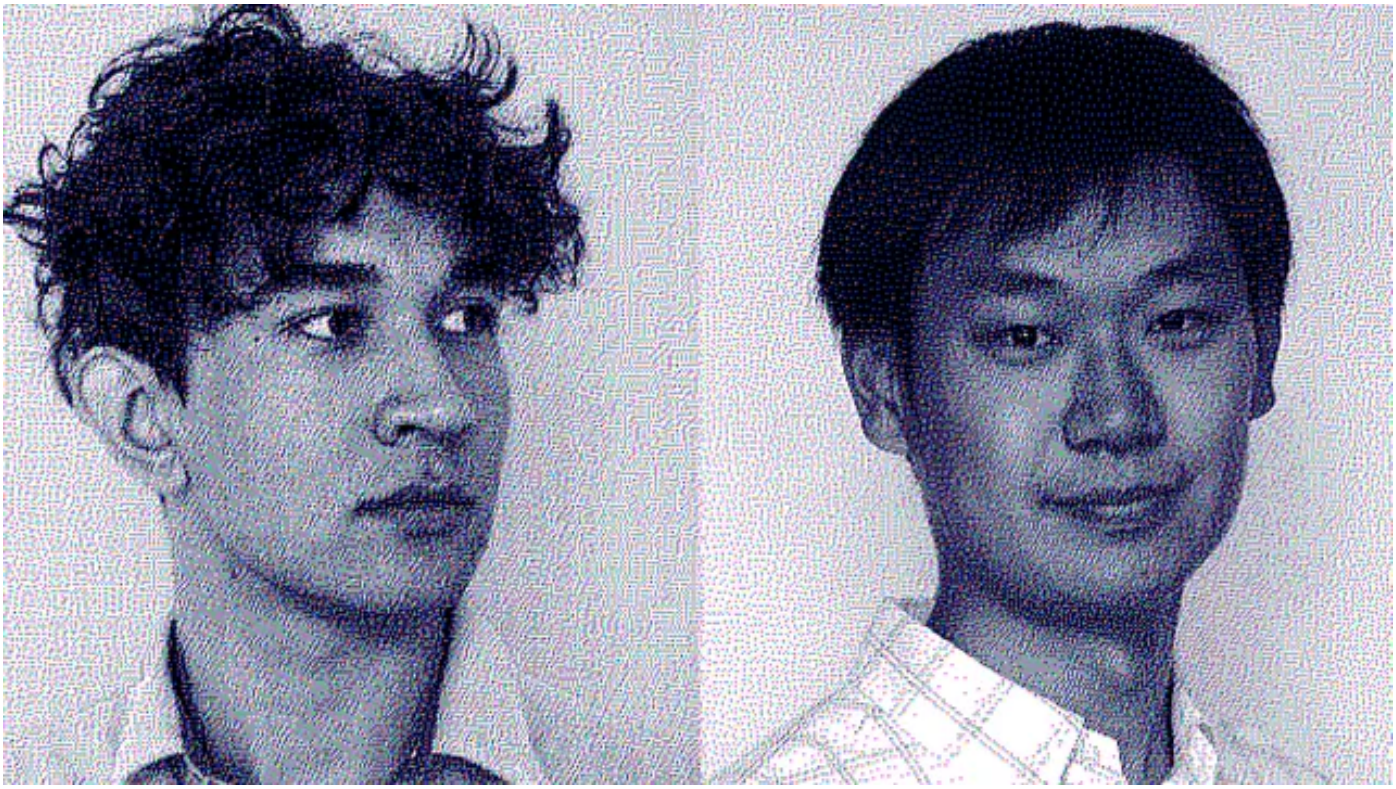
Coming from the Congo, I knew where an essential ingredient for atomic bombs was mined, even if everyone else seemed to ignore it.

NGOFEEN MPUTUBWELE

What OpenAI Really Wants

The young company sent shock waves around the world when it released ChatGPT. But that was just the start. The ultimate goal: Change everything. Yes. *Everything*.

STEVEN LEVY



The AI Detection Arms Race Is On—and College Students Are Building the Weapons

Gen Alpha is quickly developing tools that identify AI-generated text—and tools to evade detection.

CHRISTOPHER BEAM

Crispr Pioneer Jennifer Doudna Has the Guts to Take On the Microbiome

The world-famous biochemist is ready to tackle everything from immune disorders and mental illness to climate change—all by altering microbes in the digestive tract.

JENNIFER KAHN

Sundar Pichai on Google's AI, Microsoft's AI, OpenAI, and ... Did We Mention AI?

The tech giant is 25 years old. In a chatbot war. On trial for antitrust. But its CEO says Google is good for 25 more.

STEVEN LEVY